Interplay between Social Influence and Network Centrality: A Comparative Study on Shapley Centrality and Single-Node-Influence Centrality

Wei Chen
Microsoft Research
Beijing, China
weic@microsoft.com

Shang-Hua Teng
University of Southern California
Los Angeles, CA, U.S.A.
shanghua@usc.edu

ABSTRACT

We study network centrality based on dynamic influence propagation models in social networks. To illustrate our integrated mathematical-algorithmic approach for understanding the fundamental interplay between dynamic influence processes and static network structures, we focus on two basic centrality measures: (a) *Single Node Influence* (SNI) *centrality*, which measures each node's significance by its influence spread;[1] and (b) *Shapley Centrality*, which uses the Shapley value of the influence spread function — formulated based on a fundamental cooperative-game-theoretical concept — to measure the significance of nodes. We present a comprehensive comparative study of these two centrality measures. Mathematically, we present axiomatic characterizations, which precisely capture the essence of these two centrality measures and their fundamental differences. Algorithmically, we provide scalable algorithms for approximating them for a large family of social-influence instances. Empirically, we demonstrate their similarity and differences in a number of real-world social networks, as well as the efficiency of our scalable algorithms. Our results shed light on their applicability: SNI centrality is suitable for assessing individual influence in isolation while Shapley centrality assesses individuals' performance in group influence settings.

Keywords

Social network; social influence; influence diffusion model; interplay between network and influence model; network centrality; Shapley values; scalable algorithms

1. INTRODUCTION

Network science is a fast growing discipline that uses mathematical graph structures to represent real-world networks — such as the Web, Internet, social networks, biological networks, and power grids — in order to study fundamental network properties. However, network phenomena are far more complex than what can be captured only by nodes and edges, making it essential to formulate network concepts by incorporating network facets beyond graph structures [34]. For example, network centrality is a key concept in network analysis. The *centrality* of nodes, usually measured by a real-valued function, reflects their significance, importance, or crucialness within the given network. Numerous centrality measures have been proposed, based on degree, closeness, betweenness and eigenvector (i.e., PageRank) (cf. [22]). However, most of these centrality measures focus only on the static topological structures of the networks, while real-world network data include much richer interaction dynamics beyond static topology.

Influence propagation is a wonderful example of interaction dynamics in social networks. As envisioned by Domingos and Richardson [26, 14], and beautifully formulated by Kempe, Kleinberg, and Tardos [18], *social influence propagation* can be viewed as a stochastic dynamic process over an underlying static graph: After a group of nodes becomes *active*, these *seed nodes* propagate their influence through the graph structure. Even when the static graph structure of a social network is fixed, dynamic phenomena such as the spread of ideas, epidemics, and technological innovations can follow different processes. Thus, network centrality, which aims to measure nodes' importance in social influence, should be based not only on static graph structure, but also on the dynamic influence propagation process.

In this paper, we address the basic question of *how to formulate network centrality measures that reflect dynamic influence propagation*. We will focus on the study of the *interplay between social influence and network centrality*.

A social influence instance specifies a directed graph $G = (V, E)$ and an influence model $P_{\mathcal{I}}$ (see Section 2). For each $S \subseteq V$, $P_{\mathcal{I}}$ defines a stochastic influence process on G with S as the initial active set, which activates a random set $\boldsymbol{I}(S) \supseteq S$ with probability $P_{\mathcal{I}}(S, \boldsymbol{I}(S))$. Then, $\sigma(S) = \mathbb{E}[|\boldsymbol{I}(S)|]$ is the *influence spread* of S. The question above can be restated as: Given a social-influence instance $(V, E, P_{\mathcal{I}})$, how should we define the centrality of nodes in V?

A natural centrality measure for each node $v \in V$ is its influence spread $\sigma(\{v\})$. However, this measure — referred to as the *single node influence* (SNI) *centrality* — completely ignores the influence profile of groups of nodes and a node's role in such group influence. Thus, other more sensible centrality measures accounting for group influence may better capture nodes' roles in social influence. As a concrete formulation of group-influence analyses, we apply Shapley value [29] — a fundamental concept from cooperative game theory — to define a new centrality measure for social influence.

Cooperative game theory is a mathematical theory studying people's performance and behavior in coalitions (cf. [20]). Mathematically, an n-person *coalitional game* is defined by a *characteristic function* $\tau : 2^V \to \mathbb{R}$, where $V = [n]$, and $\tau(S)$ is the utility of the coalition S [29]. In this game, the *Shapley value* $\phi_v^{Shapley}(\tau)$ of $v \in V$ is v's *expected*

[1] The influence spread of a group is the expected number of nodes this group can activate as the initial active set.

marginal contribution in a random order. More precisely:

$$\phi_v^{Shapley}(\tau) = \mathbb{E}_{\boldsymbol{\pi}}[\tau(S_{\boldsymbol{\pi},v} \cup \{v\}) - \tau(S_{\boldsymbol{\pi},v})], \qquad (1)$$

where $S_{\boldsymbol{\pi},v}$ denotes the set of players preceding v in a random permutation $\boldsymbol{\pi}$ of V: The Shapley value enjoys an axiomatic characterization (see Section 2), and is widely considered to be the *fairest* measure of a player's power in a cooperative game.

Utilizing the above framework, we view influence spread $\sigma(\cdot)$ as a characteristic function, and define the *Shapley centrality* of an influence instance as the Shapley value of σ.

In this paper, we present a comprehensive comparative study of SNI and Shapley centralities. In the age of Big Data, networks are massive. Thus, an effective solution concept in network science should be both *mathematically meaningful* and *algorithmically efficient*. In our study, we will address both the conceptual and algorithmic questions.

Conceptually, influence-based centrality can be viewed as a *dimensional reduction* from the high dimensional influence model $P_{\mathcal{I}}$ to a low dimensional centrality measure. Dimensional reduction of data is a challenging task, because inevitably some information is lost. Thus, it is fundamental to characterize what each centrality measure captures.

So, "what do Shapley and SNI centralities capture? what are their basic differences?" Axiomatization is an instrumental approach for such characterization. In Section 3, we present our axiomatic characterizations. We present five axioms for Shapley centrality, and prove that it is the unique centrality measure satisfying these axioms. We do the same for the SNI centrality with three axioms. Using our axiomatic characterizations, we then provide a detailed comparison of Shapley and SNI centralities. Our characterizations show that (a) SNI centrality focuses on individual influence and would not be appropriate for models concerning group influence, such as threshold-based models. (b) Shapley centrality focuses on individuals' "irreplaceable power" in group influence settings, but may not be interpreted well if one prefer to focus on individual influence in isolation.

The computation of influence-based centralities is also a challenging problem: Exact computation of influence spread in the basic *independent cascade* and *linear-threshold* models has been shown to be #P-complete [35, 13]. Shapley centrality computation seems to be more challenging since its definition as in Eq. (1) involves various influence spreads derived from $n!$ permutations. Facing these challenges, in Section 4, we present provably-good scalable algorithms for approximating both Shapley and SNI centralities of a large family of social influence instances. Surprisingly, both algorithms share the same algorithm structure, which extends techniques from the recent algorithmic breakthroughs in influence maximization [10, 32, 31]. We further conduct empirical evaluation of Shapley and SNI centralities in a number of real-world networks. Our experiments — see Section 5 — show that our algorithms can scale up to networks with tens of millions of nodes and edges, and these two centralities are similar in several cases but also have noticeable differences.

These combined mathematical/algorithmic/empirical analyses together present (a) a systematic case study of the interplay between influence dynamics and network centrality based on Shapley and SNI centralities; (b) axiomatic characterizations for two basic centralities that precisely capture their similarities and differences; and (c) new scalable algorithms for influence models. We believe that the dual axiomatic-and-algorithmic characterization provides a comparative framework for evaluating other influence-based network concepts in the future. Due to space constraint, proofs and additional results are in [12].

1.1 Related Work

Network centrality has been extensively studied (see [22] and the references therein for a comprehensive introduction). Most classical centralities, based on degree, closeness, betweenness, eigenvector, are defined on static graphs. But some also have dynamic interpretations based on random-walks or network flows [8]. Eigenvector centrality [6] and its closely related Katz-[17] and Alpha-centrality [7] can be viewed as some forms of influence measures, since their dynamic processes are non-conservative [15], meaning that items could be replicated and propagated, similar to diffusion of ideas, opinions, etc. PageRank [11] and other random-walk related centralities correspond to conservative processes, and thus may not be suitable for propagation dynamics. Percolation centrality [25] also addresses diffusion process, but its definition only involves static percolation. None of above maps specific propagation models to network centrality. Ghosh et al. [16] maps a linear dynamic process characterized by parameterized Laplacian to centrality but the social influence models we consider in this paper are beyond such linear dynamic framework. Michalak et al. use Shapley value as network centrality [19], but they only consider five basic network games based on local sphere of influence, and their algorithms run in (least) quadratic time. To the best of our knowledge, our study is the first to explicitly map general social network influence propagation models to network centrality.

Influence propagation has been extensively studied, but most focusing on influence maximization tasks [18, 35, 13], which aims to efficiently select a set of nodes with the largest influence spread. The solution is not a centrality measure and the seeds in the solution may not be the high centrality nodes. Borgatti [9] provides clear conceptual discussions on the difference between centralities and such key player set identification problems. Algorithmically, our construction extends the idea of reverse reachable sets, recently introduced in [10, 32, 31] for scalable influence maximization.

In terms of axiomatic characterizations of network centrality, Sabidussi is the first who provides a set of axioms that a centrality measure should satisfy [27]. A number of other studies since then either provide other axioms that a centrality measure should satisfy (e.g. [23, 5, 28]) or a set of axioms that uniquely define a centrality measure (e.g. [2] on PageRank without the damping factor). All of these axiomatic characterizations focus on static graph structures, while our axiomatization focuses on the interplay between dynamic influence processes and static graph structures, and thus our study fundamentally differs from all the above characterizations. While we are heavily influenced by the axiomatic characterization of the Shapley value [29], we are also inspired by social choice theory [3], and particularly by [24] on measures of intellectual influence and [2] on PageRank.

2. INFLUENCE AND CENTRALITY

2.1 Social Influence Models

A network-influence instance is usually specified by a triple $\mathcal{I} = (V, E, P_{\mathcal{I}})$, where a directed graph $G = (V, E)$ represents the structure of a social network, and $P_{\mathcal{I}}$ defines

the influence model [18]. As an example, consider the classical discrete-time *independent cascade (IC) model*, in which each directed edge $(u, v) \in E$ has an influence probability $p_{u,v} \in [0, 1]$. At time 0, nodes in a given seed set S are activated while other nodes are inactive. At time $t \geq 1$, for any node u activated at time $t-1$, it has one chance to activate each of its inactive out-neighbor v with an independent probability $p_{u,v}$. When there is no more activation, the stochastic process ends with a random set $\boldsymbol{I}(S)$ of nodes activated during the process. The *influence spread* of S is $\sigma(S) = \mathbb{E}[|\boldsymbol{I}(S)|]$, the expected number of nodes influenced by S. Throughout the paper, we use boldface symbols to represent random variables.

Algorithmically, we will focus on the (random) *triggering model* [18], which has IC model as a special case. In this model, each $v \in V$ has a random *triggering set* $\boldsymbol{T}(v)$, drawn from a distribution defined by the influence model over the power set of all in-neighbors of v. At time $t = 0$, triggering sets $\{\boldsymbol{T}(v)\}_{v \in V}$ are drawn independently, and the seed set S is activated. At $t \geq 1$, if v is not active, it becomes activated if some $u \in \boldsymbol{T}(v)$ is activated at time $t-1$. The triggering model can be equivalently viewed under the *live-edge graph model*: (1) Draw independent random triggering sets $\{\boldsymbol{T}(v)\}_{v \in V}$; (2) form a *live-edge graph* $\boldsymbol{L} = (V, \{(u, v) : u \in \boldsymbol{T}(v)\})$, where $(u, v), u \in \boldsymbol{T}(v)$ is referred as a *live edge*. For any subgraph L of G and $S \subseteq V$, let $\Gamma(L, S)$ be the set of nodes in L reachable from set S. Then set of active nodes with seed set S is $\Gamma(\boldsymbol{L}, S)$, and influence spread $\sigma(S) = \mathbb{E}_{\boldsymbol{L}}[|\Gamma(\boldsymbol{L}, S)|] = \sum_L \Pr(\boldsymbol{L} = L) \cdot |\Gamma(L, S)|$. We say a set function $f(\cdot)$ is *monotone* if $f(S) \leq f(T)$ whenever $S \subseteq T$, and *submodular* if $f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$ whenever $S \subseteq T$ and $v \notin T$. As shown in [18], in any triggering model, $\sigma(\cdot)$ is monotone and submodular.

More generally, we define an *influence instance* as a triple $\mathcal{I} = (V, E, P_{\mathcal{I}})$, where $G = (V, E)$ represents the underlying network, and $P_{\mathcal{I}} : 2^V \times 2^V \to \mathbb{R}$ defines the probability that in the influence process, any seed set $S \subseteq V$ activates *exactly* nodes in any target set $T \subseteq V$ and no other nodes: If $\boldsymbol{I}_{\mathcal{I}}(S)$ denotes the random set activated by seed set S, then $\Pr(\boldsymbol{I}_{\mathcal{I}}(S) = T) = P_{\mathcal{I}}(S, T)$. This probability profile is commonly defined by a succinct influence model, such as the triggering model, which interacts with network G. We also require that: (a) $P_{\mathcal{I}}(\emptyset, \emptyset) = 1$, $P_{\mathcal{I}}(\emptyset, T) = 0$, $\forall T \neq \emptyset$, and (b) if $S \not\subseteq T$ then $P_{\mathcal{I}}(S, T) = 0$, i.e., S always activates itself ($S \subseteq \boldsymbol{I}_{\mathcal{I}}(S)$). Such model is also referred to as the *progressive* influence model. The *influence spread* of S is:

$$\sigma_{\mathcal{I}}(S) = \mathbb{E}[|\boldsymbol{I}_{\mathcal{I}}(S)|] = \sum_{T \subseteq V, S \subseteq T} P_{\mathcal{I}}(S, T) \cdot |T|.$$

2.2 Coalitional Games and Shapley Values

An n-person *coalitional game* over $V = [n]$ is specified by a *characteristic function* $\tau : 2^V \to \mathbb{R}$, where for any coalition $S \subseteq V$, $\tau(S)$ denotes the *cooperative utility* of S. In cooperative game theory, a *ranking function* ϕ is a mapping from a characteristic function τ to a vector in $\mathbb{R}^n$. A fundamental solution concept of cooperative game theory is the ranking function given by the *Shapley value* [29]: Let Π be the set of all permutations of V. For any $v \in V$ and $\pi \in \Pi$, let $S_{\pi,v}$ denote the set of nodes in V preceding v in permutation π. Then, $\forall v \in V$:

$$\begin{aligned}\phi_v^{Shapley}(\tau) =& \frac{1}{n!} \sum_{\pi \in \Pi} (\tau(S_{\pi,v} \cup \{v\}) - \tau(S_{\pi,v})) \\ =& \sum_{S \subseteq V \setminus \{v\}} \frac{|S|!(n - |S| - 1)!}{n!} (\tau(S \cup \{v\}) - \tau(S)).\end{aligned}$$

We use $\boldsymbol{\pi} \sim \Pi$ to denote that $\boldsymbol{\pi}$ is a random permutation uniformly drawn from Π. Then:

$$\phi_v^{Shapley}(\tau) = \mathbb{E}_{\boldsymbol{\pi} \sim \Pi}[\tau(S_{\boldsymbol{\pi},v} \cup \{v\}) - \tau(S_{\boldsymbol{\pi},v})]. \quad (2)$$

The Shapley value of v measures v's marginal contribution over the set preceding v in a random permutation.

Shapley [29] proved a remarkable representation theorem: The Shapley value is the unique ranking function that satisfies all the following four conditions: (1) Efficiency: $\sum_{v \in V} \phi_v(\tau) = \tau(V)$. (2) Symmetry: For any $u, v \in V$, if $\tau(S \cup \{u\}) = \tau(S \cup \{v\}), \forall S \subseteq V \setminus \{u, v\}$, then $\phi_u(\tau) = \phi_v(\tau)$. (3) Linearity: For any two characteristic functions τ and ω, for any $\alpha, \beta > 0$, $\phi(\alpha\tau + \beta\omega) = \alpha\phi(\tau) + \beta\phi(\omega)$. (4) Null Player: For any $v \in V$, if $\tau(S \cup \{v\}) - \tau(S) = 0, \forall S \subseteq V \setminus \{v\}$, then $\phi_v(\tau) = 0$. Efficiency states that the total utility is fully distributed. Symmetry states that two players' ranking values should be the same if they have the identical marginal utility profile. Linearity states that the ranking values of the weighted sum of two coalitional games is the same as the weighted sum of their ranking values. Null Player states that a player's ranking value should be zero if the player has zero marginal utility to every subset.

2.3 Shapley and SNI Centrality

The influence-based centrality measure aims at assigning a value for every node under every influence instance:

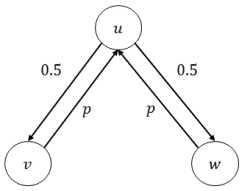
Definition 1 (Centrality Measure). *An* (influence-based) centrality measure ψ *is a mapping from an influence instance* $\mathcal{I} = (V, E, P_{\mathcal{I}})$ *to a real vector* $(\psi_v(\mathcal{I}))_{v \in V} \in \mathbb{R}^{|V|}$.

The *single node influence* (SNI) *centrality*, denoted by $\psi_v^{SNI}(\mathcal{I})$, assigns the influence spread of node v as v's centrality measure: $\psi_v^{SNI}(\mathcal{I}) = \sigma_{\mathcal{I}}(\{v\})$.

The *Shapley centrality*, denoted by $\psi^{Shapley}(\mathcal{I})$, is the Shapley value of the influence spread function $\sigma_{\mathcal{I}}$: $\psi^{Shapley}(\mathcal{I}) = \phi^{Shapley}(\sigma_{\mathcal{I}})$. As a subtle point, note that $\phi^{Shapley}$ maps from a $2^{|V|}$ dimensional τ to a $|V|$-dimensional vector, while, formally, $\psi^{Shapley}$ maps from $P_{\mathcal{I}}$ — whose dimensions is close to $2^{2|V|}$ — to a $|V|$-dimensional vector.

To help understand these definitions, Figure 1 provides a simple example of a 3-node graph in the IC model with influence probabilities shown on the edges. The associated table shows the result for Shapley and SNI centralities. While SNI is straightforward in this case, the Shapley centrality calculation already looks complex. Due to space constraint, we left readers to verify the computation. Based on the result, we find that for interval $p \in (1/2, 2/3)$, Shapley and SNI centralities do not align in ranking: Shapley places v, w higher than u while SNI puts u higher than v, w. This simple example already illustrates that (a) computing Shapley centrality could be a nontrivial task; and (b) the relationship between Shapley and SNI centralities could be complicated. Addressing both the computation and characterization questions are the subject of the remaining sections.

3. AXIOMATIC CHARACTERIZATION



	Shapley	SNI
u	$\frac{1}{3}\cdot 2+\frac{1}{3}(1-p)\cdot 1.5+\frac{1}{3}(1-p)^2$ $=\frac{1}{6}(9-7p+2p^2)$	2
v,w	$\frac{1}{3}(1+1.5p)+\frac{1}{6}\cdot 0.5+$ $\frac{1}{6}((1-p)(1+p)+0.5p)+\frac{1}{3}\cdot 0.5$ $=\frac{1}{12}(9+7p-2p^2)$	$1+1.5p$

Figure 1: Example on Shapley and SNI centrality.

In this section, we present two sets of axioms uniquely characterizing Shapley and SNI centralities, respectively, based on which we analyze their similarities and differences.

3.1 Axioms for Shapley Centrality

Our set of axioms for characterizing the Shapley centrality is adapted from the classical Shapley's axioms [29].

The first axiom states that labels on the nodes should have no effect on centrality measures. This ubiquitous axiom is similar to the isomorphic axiom in some other centrality characterizations, e.g. [27].

Axiom 1 (Anonymity). *For any influence instance* $\mathcal{I}=(V,E,P_{\mathcal{I}})$, *and permutation* $\pi\in\Pi$, $\psi_v(\mathcal{I})=\psi_{\pi(v)}(\pi(\mathcal{I}))$, $\forall v\in V$.

In Axiom 1, $\pi(\mathcal{I})=(\pi(V),\pi(E),P_{\pi(\mathcal{I})})$ denotes the isomorphic instance: (1) $\forall u,v\in V$, $(\pi(u),\pi(v))\in\pi(E)$ iff $(u,v)\in E$, and (2) $\forall S,T\subseteq V$, $P_{\mathcal{I}}(S,T)=P_{\pi(\mathcal{I})}(\pi(S),\pi(T))$.

The second axiom states that the centrality measure divides the total share of influence $|V|$. In other words, the average centrality is normalized to 1.

Axiom 2 (Normalization). *For every influence instance* $\mathcal{I}=(V,E,P_{\mathcal{I}})$, $\sum_{v\in V}\psi_v(\mathcal{I})=|V|$.

The next axiom characterizes the centrality of a type of extreme nodes in social influence. In instance $\mathcal{I}=(V,E,P_{\mathcal{I}})$, we say $v\in V$ is a *sink node* if $\forall S,T\subseteq V\setminus\{v\}$, $P_{\mathcal{I}}(S\cup\{v\},T\cup\{v\})=P_{\mathcal{I}}(S,T)+P_{\mathcal{I}}(S,T\cup\{v\})$. In the extreme case when $S=T=\emptyset$, $P_{\mathcal{I}}(\{v\},\{v\})=1$, i.e., v can only influence itself. When v joins another S to form a seed set, the influence to a target $T\cup\{v\}$ can always be achieved by S alone (except perhaps the influence to v itself). In the triggering model, a sink node is (indeed) a node without outgoing edges, matching the name "sink".

Because a sink node v has no influence on other nodes, we can "remove" it and obtain a projection of the influence model on the network without v: Let $\mathcal{I}\setminus\{v\}=(V\setminus\{v\},E\setminus\{v\},P_{\mathcal{I}\setminus\{v\}})$ denote the *projected* instance over $V\setminus\{v\}$, where $E\setminus\{v\}=\{(i,j)\in E:v\notin\{i,j\}\}$ and $P_{\mathcal{I}\setminus\{v\}}$ is the influence model such that for all $S,T\subseteq V\setminus\{v\}$:

$$P_{\mathcal{I}\setminus\{v\}}(S,T)=P_{\mathcal{I}}(S,T)+P_{\mathcal{I}}(S,T\cup\{v\}).$$

Intuitively, since sink node v is removed, the previously distributed influence from S to T and $T\cup\{v\}$ is merged into the influence from S to T in the projected instance. For the triggering model, influence projection is simply removing the sink node v and its incident incoming edges without changing the triggering set distribution of any other nodes.

Axiom 3 below considers the simple case when the influence instance has two sink nodes $u,v\in V$. In such a case, u and v have no influence to each other, and they influence no one else. Thus, their centrality should be fully determined by $V\setminus\{u,v\}$: Removing one sink node — say v — should not affect the centrality measure of another sink node u.

Axiom 3 (Independence of Sink Nodes). *For any influence instance* $\mathcal{I}=(V,E,P_{\mathcal{I}})$, *for any pair of sink nodes* $u,v\in V$ *in* $\mathcal{I}$, *it should be the case:* $\psi_u(\mathcal{I})=\psi_u(\mathcal{I}\setminus\{v\})$.

The next axiom considers *Bayesian social influence* through a given network: Given a graph $G=(V,E)$, and r influence instances on G: $\mathcal{I}^\eta=(V,E,P_{\mathcal{I}^\eta})$ with $\eta\in[r]$. Let $\lambda=(\lambda_1,\lambda_2,\dots,\lambda_r)$ be a prior distribution on $[r]$, i.e. $\sum_{\eta=1}^r\lambda_\eta=1$, and $\lambda_\eta\geq 0$, $\forall\eta\in[r]$. The *Bayesian influence instance* $\mathcal{I}_{\mathcal{B}(\{\mathcal{I}^\eta\},\lambda)}$ has the following influence process for a seed set $S\subseteq V$: (1) Draw a random index $\boldsymbol{\eta}\in[r]$ according to distribution λ (denoted as $\boldsymbol{\eta}\sim\lambda$). (2) Apply the influence process of $\mathcal{I}^{\boldsymbol{\eta}}$ with seed set S to obtain the activated set T. Equivalently, we have for all $S,T\subseteq V$, $P_{\mathcal{I}_{\mathcal{B}(\{\mathcal{I}^\eta\},\lambda)}}(S,T)=\sum_{\eta=1}^r\lambda_\eta P_{\mathcal{I}^\eta}(S,T)$. In the triggering model, we can view each live-edge graph and the deterministic diffusion on it via reachability as an influence instance, and the diffusion of the triggering model is by the Bayesian (or convex) combination of these live-edge instances. The next axiom reflects the linearity-of-expectation principle:

Axiom 4 (Bayesian Influence). *For any network* $G=(V,E)$ *and Bayesian social-influence model* $\mathcal{I}_{\mathcal{B}(\{\mathcal{I}^\eta\},\lambda)}$:

$$\psi_v(\mathcal{I}_{\mathcal{B}(\{\mathcal{I}^\eta\},\lambda)})=\mathbb{E}_{\boldsymbol{\eta}\sim\lambda}\left[\psi_v(\mathcal{I}^{\boldsymbol{\eta}})\right]=\sum_{\eta=1}^r\lambda_\eta\cdot\psi_v(\mathcal{I}^\eta),\forall v\in V.$$

The above axiom essentially says that the centrality of a Bayesian instance before realizing the actual model $\mathcal{I}^\eta$ is the same as the expected centrality after realizing $\mathcal{I}^\eta$.

The last axiom characterizes the centrality of a family of simple social-influence instances. For any $\emptyset\subset R\subseteq U\subseteq V$, a *critical set instance* $\mathcal{I}_{R,U}=(V,E,P_{\mathcal{I}_{R,U}})$ is such that: (1) The network $G=(V,E)$ contains a complete directed bipartite sub-graph from R to $U\setminus R$, together with isolated nodes $V\setminus U$. (2) For all $S\supseteq R$, $P_{\mathcal{I}_{R,U}}(S,U\cup S)=1$, and (3) For all $S\not\supseteq R$, $P_{\mathcal{I}_{R,U}}(S,S)=1$. In $\mathcal{I}_{R,U}$, R is called the *critical set*, and U is called the *target set*. In other words, a seed set containing R activates all nodes in U, but missing any node in R the seed set only activates itself. We use $\mathcal{I}_{R,v}$ to denote the special case of $U=R\cup\{v\}$ and $V=U$. That is, only if all nodes in R work together they can activate v.

Axiom 5 (Bargaining with Critical Sets). *In any critical set instance* $\mathcal{I}_{R,v}$, *the centrality of* v *is* $\frac{|R|}{|R|+1}$, *i.e.* $\psi_v(\mathcal{I}_{R,v})=\frac{|R|}{|R|+1}$.

Qualitatively, Axiom 5 together with Normalization and Anonymity axioms implies that the relative importance of v comparing to a node in the critial set R increases when $|R|$ increases, which is reasonable because when the critical set R grows, individuals in R becomes weaker and v becomes relatively stronger. The actual quantity can be explained by Nash's solution to the bargaining game [21] (see [12]).

Our first axiomatic representation theorem can now be stated as the following:

Theorem 1. (Axiomatic Characterization of Shapley Centrality) *The Shapley centrality* $\psi^{Shapley}$ *is the unique centrality measure that satisfies Axioms 1-5. Moreover, every axiom in this set is independent of others.*

The soundness of this representation theorem — that the Shapley centrality satisfies all axioms — is relatively simple. However, because of the intrinsic complexity in influence models, the uniqueness proof is in fact complex. We give a high-level proof sketch here and the full proof is in [12]. We follow Myerson's proof strategy [20] of Shapley's theorem. The probabilistic profile $P_{\mathcal{I}}$ of influence instance $\mathcal{I} = (V, E, P_{\mathcal{I}})$ is viewed as a vector in a large space R^M, where M is the number of independent dimensions in $P_{\mathcal{I}}$. Bayesian Influence Axiom enforces that any conforming centrality measure is an affine mapping from R^M to $\mathbb{R}^n$. We then prove that the critical set instances $\mathcal{I}_{R,U}$ form a full-rank basis of the linear space R^M. Finally, we prove that any axiom-conforming centrality measure over critical set instances (and the additional null instance in which every node is a sink node) must be unique. The uniqueness of the critical set instances and the null instance, the linear independence of critical set instances in R^M, plus the affine mapping from R^M to $\mathbb{R}^n$, together imply that the centrality measure of every influence instance is uniquely determined. Our overall proof is more complex and — to a certain degree — more subtle than Myerson's proof, because our axiomatic framework is based on the influence model in a much larger dimensional space compared to the subset utility functions. Finally, for independence, we need to show that for each axiom, we can construct an alternative centrality measure if the axiom is removed. Except for Axiom 5, the constructions and the proofs for other axioms are nontrivial, and they shed lights on how related centrality measures could be formed when some conditions are relaxed.

3.2 Axioms for SNI Centrality

We first examine which of Axioms 1-5 are satisfied by SNI centrality. It is easy to verify that Anonymity and Bayesian Influence Axioms hold for SNI centrality. For the Independence of Sink Node Axiom (Axiom 3), since every sink node can only influence itself, its SNI centrality is 1. Thus, Axiom 3 is satisfied by SNI because of a stronger reason.

For the Normalization Axiom (Axiom 2), the sum of single node influence is typically more than the total number of nodes (e.g., when the influence spread is submodular), and thus Axiom 2 does not hold for SNI centrality. The Bargaining with Critical Sets Axiom (Axiom 5) does not hold either, since node v in $\mathcal{I}_{R,v}$ is a sink node and thus its SNI centrality is 1.

We now present our axiomatic characterization of SNI centrality, which will retain Bayesian Influence Axiom 4, strengthen Independence of Sink Node Axiom 3, and recharacterize the centrality of a node in a critical set:

Axiom 6 (Uniform Sink Nodes). *Every sink node has centrality* 1.

Axiom 7 (Critical Nodes). *In any critical set instance $\mathcal{I}_{R,U}$, the centrality of a node $w \in R$ is* 1 *if* $|R| > 1$, *and is* $|U|$ *if* $|R| = 1$.

These three axioms are sufficient to uniquely characterize SNI centrality, as they also imply Anonymity Axiom:

Theorem 2. (Axiomatic Characterization of SNI Centrality) *The SNI centrality ψ^{SNI} is the unique centrality measure that satisfies Axioms 4, 6, and 7. Moreover, each of these axioms is independent of the others.*

Theorems 1 and 2 establish the following appealing property: Even though all our axioms are on probabilistic profiles $P_{\mathcal{I}}$ of influence instances, the unique centrality measure satisfying these axioms is in fact fully determined by the influence spread profile $\sigma_{\mathcal{I}}$. We find this amazing because the distribution profile $P_{\mathcal{I}}$ has much higher dimensionality than its influence-spread profile $\sigma_{\mathcal{I}}$.

3.3 Shapley Centrality versus SNI Centrality

We now provide a comparative analysis between Shapley and SNI centralities based on their definitions, axiomatic characterizations, and various other properties they satisfy.

Comparison by definition. The definition of SNI centrality is more straightforward as it uses individual node's influence spread as the centrality measure. Shapley centrality is more sophisticatedly formulated, involving groups' influence spreads. SNI centrality disregards the influence profile of groups. Thus, it may limit its usage in more complex situations where group influences should be considered. Meanwhile, Shapley centrality considers group influence in a particular way involving marginal influence of a node on a given group randomly ordered before the node. Thus, Shapley centrality is more suitable for assessing marginal influence of a node in a group setting.

Comparison by axiomatic characterization. Both SNI and Shapley centralities satisfy Anonymity, Independence of Sink Nodes, and Bayesian Influence axioms, which seem to be natural axioms for desirable social-influence centrality measures. Their unique axioms characterize exactly their differences. The first difference is on the Normalization Axiom, satisfied by Shapley but not SNI centrality. This indicates that Shapley centrality aims at dividing the total share of possible influence spread $|V|$ among all nodes, but SNI centrality does not enforce such share division among nodes. If we artificially normalize the SNI centrality values of all nodes to satisfy the Normalization Axiom, the normalized SNI centrality would not satisfy the Bayesian Influence Axiom. (In fact, it is not easy to find a new characterization for the normalized SNI centrality similar to Theorem 2.) We will see shortly that the Normalization Axiom would also cause a drastic difference between the two centrality measures for the symmetric IC influence model.

The second difference is on their treatment of sink nodes, exemplified by sink nodes in the critical set instances. For SNI centrality, sink nodes are always treated with the same centrality of 1 (Axiom 6). But the Shapley centrality of a sink node may be affected by other nodes that influence the sink. In particular, for the critical set instance $\mathcal{I}_{R,v}$, v has centrality $|R|/(|R|+1)$, which increases with R. As discussed earlier, larger R indicates v is getting stronger comparing to nodes in R. In this aspect, Shapley centrality assignment is sensible. Overall, when considering v's centrality, SNI centrality disregards other nodes' influence to v while Shapley centrality considers other nodes' influence to v.

The third difference is their treatment of critical nodes in the critical set instances. For SNI centrality, in the critical set instance $\mathcal{I}_{R,v}$, Axiom 7 obliviously assigns the same value 1 for nodes $u \in R$ whenever $|R| > 1$, effectively equalizing the centrality of node $u \in R$ with v. In contrast, Shapley centrality would assign $u \in R$ a value of $1 + \frac{1}{|R|(|R|+1)}$, decreasing with R but is always larger than v's centrality of $\frac{|R|}{|R|+1}$. Thus Shapley centrality assigns more sensible values

in this case, because $u \in R$ as part of a coalition should have larger centrality than v, who has no influence power at all. We believe this shows the limitation of the SNI centrality — it only considers individual influence and disregards group influence. Since the critical set instances reflect the threshold behavior in influence propagation — a node would be influenced only after the number of its influenced neighbors reach certain threshold — this suggests that SNI centrality could be problematic in threshold-based influence models.

Comparison by additional properties. Finally, we compare additional properties they satisfy. First, it is straightforward to verify that both centrality measures satisfy the *Independence of Irrelevant Alternatives (IIA))* property: If an instance $\mathcal{I} = (V, E, P_{\mathcal{I}})$ is the *union of two disjoint and independent influence instances*, $\mathcal{I}_1 = (V_1, E_1, P_{\mathcal{I}_1})$ and $\mathcal{I}_2 = (V_2, E_2, P_{\mathcal{I}_2})$, then for $k \in \{1, 2\}$ and any $v \in V_k$: $\psi_v(\mathcal{I}) = \psi_v(\mathcal{I}_k)$.

The IIA property together with the Normalization Axiom leads to a clear difference between SNI and Shapley centrality. Consider an example of two undirected and connected graphs G_1 with 10 nodes and G_2 with 3 nodes, and the IC model on them with edge probability 1. Both SNI and Shapley centralities assign same values to nodes within each graph, but due to normalization, Shapley assigns 1 to all nodes, while SNI assigns 10 to nodes in G_1 and 3 to nodes in G_2. The IIA property ensures that the centrality does not change when we put G_1 and G_2 together. That is, SNI considers nodes in G_1 more important while Shapley considers them the same. While SNI centrality makes sense from individual influence point of view, the view of Shapley centrality is that a node in G_1 is easily replaceable by any of the other 9 nodes in G_1 but a node in G_2 is only replaceable by two other nodes in G_2. Shapley centrality uses marginal influence in randomly ordered groups to determine that the "replaceability factor" cancels out individual influence and assigns same centrality to all nodes.

The above example generalizes to the symmetric IC model where $p_{u,v} = p_{v,u}$, $\forall u, v \in V$: *Every node has Shapley centrality of* 1 *in such models.* The technical reason is that such models have an equivalent *undirected* live-edge graph representation, containing a number of connected components just like the above example. The Shapley symmetry in the symmetric IC model may sound counter-intuitive, since it appears to be independent of network structures or edge probability values. But we believe what it unveils is that symmetric IC model might be an unrealistic model in practice — it is hard to imagine that between every pair of individuals the influence strength is symmetric. For example, in a star graph, when we perceive that the node in the center has higher centrality, it is not just because of its center position, but also because that it typically exerts higher influence to its neighbors than the reverse direction. This exactly reflects our original motivation that mere positions in a static network may not be an important factor in determining the node centrality, and what important is the effect of individual nodes participating in the dynamic influence process.

From the above discussions, we clearly see that (a) SNI centrality focuses on individual influence in isolation, while (b) Shapley centrality focuses on marginal influence in group influence settings, and measures the *irreplaceability* of the nodes in some sense.

4. SCALABLE ALGORITHMS

In this section, we first give a sampling-based algorithm for approximating the Shapley centrality $\psi^{Shapley}(\mathcal{I})$ of any influence instance in the triggering model. We then give a slight adaptation to approximate SNI centrality. In both cases, we characterize the performance of our algorithms and prove that they are scalable for a large family of social-influence instances. In next section, we empirically show that these algorithms are efficient for real-world networks.

4.1 Algorithm for Shapley Centrality

In this subsection, we use ψ as a shorthand for $\psi^{Shapley}$. Let $n = |V|$ and $m = |E|$. To precisely state our result, we make the following general computational assumption, as in [32, 31]:

ASSUMPTION 1. *The time to draw a random triggering set $\boldsymbol{T}(v)$ is proportional to the in-degree of v.*

The key combinatorial structures that we use are the following random sets generated by the *reversed diffusion process* of the triggering model. A *(random) reverse reachable (RR) set* $\boldsymbol{R}$ is generated as follows: (0) Initially, $\boldsymbol{R} = \emptyset$. (1) Select a node $\boldsymbol{v} \sim V$ uniformly at random (called the *root* of $\boldsymbol{R}$), and add $\boldsymbol{v}$ to $\boldsymbol{R}$. (2) Repeat the following process until every node in $\boldsymbol{R}$ has a triggering set: For every $\boldsymbol{u} \in \boldsymbol{R}$ not yet having a triggering set, draw its random triggering set $\boldsymbol{T}(\boldsymbol{u})$, and add $\boldsymbol{T}(\boldsymbol{u})$ to $\boldsymbol{R}$. Suppose $\boldsymbol{v} \sim V$ is selected in Step (1). The reversed diffusion process uses $\boldsymbol{v}$ as the seed, and follows the incoming edges instead of the outgoing edges to iteratively "influence" triggering sets. Equivalently, an RR set $\boldsymbol{R}$ is the set of nodes in a random live-edge graph $\boldsymbol{L}$ that can reach node $\boldsymbol{v}$.

The following key lemma elegantly connects RR sets with Shapley centrality. We will defer its intuitive explanation to the end of this section. Let $\boldsymbol{\pi}$ be a random permutation on V. Let $\mathbb{I}\{\mathcal{E}\}$ be the indicator function for event $\mathcal{E}$.

LEMMA 1 (SHAPLEY CENTRALITY IDENTITY). *Let $\boldsymbol{R}$ be a random RR set. Then, $\forall u \in V$, u's Shapley centrality is $\psi_u = n \cdot \mathbb{E}_{\boldsymbol{R}}[\mathbb{I}\{u \in \boldsymbol{R}\}/|\boldsymbol{R}|]$.*

This lemma is instrumental to our scalable algorithm. It guarantees that we can use random RR sets to build *unbiased estimators* of Shapley centrality. Our algorithm ASV-RR (standing for "Approximate Shapley Value by RR Set") is presented in Algorithm 1. It takes ε, ℓ, and k as input parameters, representing the relative error, the confidence of the error, and the number of nodes with top Shapley values that achieve the error bound, respectively. Their exact meaning will be made clear in Theorem 3.

ASV-RR follows the structure of the IMM algorithm of [31] but with some key differences. In Phase 1, Algorithm 1 estimates the number of RR sets needed for the Shapley estimator. For a given parameter k, we first estimate a lower bound $\boldsymbol{LB}$ of the k-th largest Shapley centrality $\psi^{(k)}$. Following a similar structure as the sampling method in IMM [31], the search of the lower bound is carried out in at most $\lfloor \log_2 n \rfloor - 1$ iterations, each of which halves the lower bound target $x = n/2^i$ and obtains the number of RR sets θ_i needed in this iteration (line 6). The key difference is that we do not need to store the RR sets and compute a max cover. Instead, for every RR set $\boldsymbol{R}$, we only update the estimate $\boldsymbol{est_u}$ of each node $\boldsymbol{u} \in \boldsymbol{R}$ with an additional $1/|\boldsymbol{R}|$ (line 9), which is based on Lemma 1. In each iteration, we select the k-th largest estimate (line 11) and plug it into the condition

Input: Network: $G = (V, E)$; Parameters: random triggering set distribution $\{\boldsymbol{T}(v)\}_{v\in V}$, $\varepsilon > 0$, $\ell > 0$, $k \in [n]$
Output: $\hat{\boldsymbol{\psi}}_v$, $\forall v \in V$: estimated centrality measure

1: {Phase 1. Estimate the number of RR sets needed }
2: $\boldsymbol{LB} = 1$; $\varepsilon' = \sqrt{2}\cdot\varepsilon$; $\theta_0 = 0$
3: $\boldsymbol{est}_v = 0$ for every $v \in V$
4: **for** $i = 1$ to $\lfloor \log_2 n \rfloor - 1$ **do**
5: $\quad x = n/2^i$
6: $\quad \theta_i = \left\lceil \frac{n\cdot((\ell+1)\ln n + \ln\log_2 n + \ln 2)\cdot(2+\frac{2}{3}\varepsilon')}{\varepsilon'^2\cdot x} \right\rceil$
7: $\quad$ **for** $j = 1$ to $\theta_i - \theta_{i-1}$ **do**
8: $\quad\quad$ generate a random RR set $\boldsymbol{R}$
9: $\quad\quad$ for every $\boldsymbol{u} \in \boldsymbol{R}$, $\boldsymbol{est_u} = \boldsymbol{est_u} + 1/|\boldsymbol{R}|$
10: $\quad$ **end for**
11: $\quad \boldsymbol{est}^{(k)}$ = the k-th largest value in $\{\boldsymbol{est}_v\}_{v\in V}$
12: $\quad$ **if** $n\cdot \boldsymbol{est}^{(k)}/\theta_i \geq (1+\varepsilon')\cdot x$ **then**
13: $\quad\quad \boldsymbol{LB} = n\cdot \boldsymbol{est}^{(k)}/(\theta_i\cdot(1+\varepsilon'))$
14: $\quad\quad$ **break**
15: $\quad$ **end if**
16: **end for**
17: $\boldsymbol{\theta} = \left\lceil \frac{n((\ell+1)\ln n + \ln 4)(2+\frac{2}{3}\varepsilon)}{\varepsilon^2\cdot \boldsymbol{LB}} \right\rceil$
18: {Phase 2. Estimate Shapley value}
19: $\boldsymbol{est}_v = 0$ for every $v \in V$
20: **for** $j = 1$ to $\boldsymbol{\theta}$ **do**
21: $\quad$ generate a random RR set $\boldsymbol{R}$
22: $\quad$ for every $\boldsymbol{u} \in \boldsymbol{R}$, $\boldsymbol{est_u} = \boldsymbol{est_u} + 1/|\boldsymbol{R}|$
23: **end for**
24: for every $v \in V$, $\hat{\boldsymbol{\psi}}_v = n\cdot \boldsymbol{est}_v/\boldsymbol{\theta}$
25: return $\hat{\boldsymbol{\psi}}_v$, $v \in V$

Algorithm 1: ASV-RR$(G, \boldsymbol{T}, \varepsilon, \ell, k)$

in line 12. Once the condition holds, we calculate the lower bound $\boldsymbol{LB}$ in line 13 and break the loop. Next we use this $\boldsymbol{LB}$ to obtain the number of RR sets $\boldsymbol{\theta}$ needed in Phase 2 (line 17). In Phase 2, we first reset the estimates (line 19), then generate $\boldsymbol{\theta}$ RR sets and again updating $\boldsymbol{est}_u$ with $1/|\boldsymbol{R}|$ increment for each $\boldsymbol{u} \in \boldsymbol{R}$ (line 22). Finally, these estimates are transformed into the Shapley estimation in line 24.

Unlike IMM, we do not reuse the RR sets generated in Phase 1, because it would make the RR sets dependent and the resulting Shapley centrality estimates biased. Moreover, our entire algorithm does not need to store any RR sets, and thus ASV-RR does not have the memory bottleneck encountered by IMM when dealing with large networks. The following theorem summarizes the performance of Algorithm 1, where ψ and $\psi^{(k)}$ are Shapley centrality and k-th largest Shapley centrality value, respectively.

THEOREM 3. *For any $\epsilon > 0$, $\ell > 0$, and $k \in [n]$, Algorithm* ASV-RR *returns an estimated Shapley value $\hat{\boldsymbol{\psi}}_v$ that satisfies (a) unbiasedness: $\mathbb{E}[\hat{\boldsymbol{\psi}}_v] = \psi_v, \forall v \in V$; (b) absolute normalization: $\sum_{v\in V}\hat{\boldsymbol{\psi}}_v = n$ in every run; and (c) robustness: under the condition that $\psi^{(k)} \geq 1$, with probability at least $1 - \frac{1}{n^\ell}$:*

$$\begin{cases} |\hat{\boldsymbol{\psi}}_v - \psi_v| \leq \varepsilon\psi_v & \forall v \in V \text{ with } \psi_v > \psi^{(k)}, \\ |\hat{\boldsymbol{\psi}}_v - \psi_v| \leq \varepsilon\psi^{(k)} & \forall v \in V \text{ with } \psi_v \leq \psi^{(k)}. \end{cases} \tag{3}$$

Under Assumption 1 and the condition $\ell \geq (\log_2 k - \log_2\log_2 n)/\log_2 n$, the expected running time of ASV-RR *is $O(\ell(m+n)\log n\cdot\mathbb{E}[\sigma(\tilde{\boldsymbol{v}})]/(\psi^{(k)}\varepsilon^2))$, where $\mathbb{E}[\sigma(\tilde{\boldsymbol{v}})]$ is the expected influence spread of a random node $\tilde{\boldsymbol{v}}$ drawn from V with probability proportional to the in-degree of $\tilde{\boldsymbol{v}}$.*

Eq. (3) above shows that for the top k Shapley values, ASV-RR guarantees the multiplicative error of ε relative to node's own Shapley value, and for the rest Shapley value, the error is relative to the k-th largest Shapley value $\psi^{(k)}$. This is reasonable since typically we only concern nodes with top Shapley values. For time complexity, the condition $\ell \geq (\log_2 k - \log_2\log_2 n)/\log_2 n$ always hold if $k \leq \log_2 n$ or $\ell \geq 1$. When fixing ε as a constant, the running time depends almost linearly on the graph size $(m+n)$ multiplied by a ratio $\mathbb{E}[\sigma(\tilde{\boldsymbol{v}})]/\psi^{(k)}$. This ratio is upper bounded by the ratio between the largest single node influence and the k-th largest Shapley value. When these two quantities are about the same order, we have a near-linear time, i.e., scalable [33], algorithm. Our experiments show that in most datasets tested the ratio $\mathbb{E}[\sigma(\tilde{\boldsymbol{v}})]/\psi^{(k)}$ is indeed less than 1. Moreover, if we could relax the robustness requirement in Eq. (3) to allow the error of $|\hat{\boldsymbol{\psi}}_v - \psi_v|$ to be relative to the largest single node influence, then we could indeed slightly modify the algorithm to obtain a near-linear-time algorithm without the ratio $\mathbb{E}[\sigma(\tilde{\boldsymbol{v}})]/\psi^{(k)}$ in the time complexity (see [12]).

The accuracy of ASV-RR is based on Lemma 1 while the time complexity analysis follows a similar structure as in [31]. Due to space limit, the proofs of Lemma 1 and Theorems 3 are presented in our full report [12]. Here, we give a high-level explanation. In the triggering model, as for influence maximization [10, 32, 31], a random RR set $\boldsymbol{R}$ can be equivalently obtained by first generating a random live-edge graph $\boldsymbol{L}$, and then constructing $\boldsymbol{R}$ as the set of nodes that can reach a random $\boldsymbol{v} \sim V$ in $\boldsymbol{L}$. The fundamental equation associated with this live-edge graph process is:

$$\sigma(S) = \sum_L \Pr_{\boldsymbol{L}}(\boldsymbol{L} = L)\Pr_{\boldsymbol{v}}(\boldsymbol{v} \in \Gamma(L, S))\cdot n. \tag{4}$$

Our Lemma 1 is the result of the following crucial observations: First, the Shapley centrality ψ_u of node $u \in V$ can be equivalently formulated as the expected Shapley centrality of u over all live-edge graphs and random choices of root $\boldsymbol{v}$, from Eq. (4). The chief advantage of this formulation is that it localizes the contribution of marginal influences: On a fixed live-graph L and root $v \in V$, we only need to compute the marginal influence of u in terms of activating v to obtain the Shapley contribution of the pair. We do not need to compute the marginal influences of u for activating other nodes. Lemma 1 then follows from our second crucial observation. When R is the fixed set that can reach v in L, the marginal influence of u activating v in a random order is 1 if and only if the following two conditions hold concurrently: (a) u is in R — so u has chance to activate v, and (b) u is ordered before any other node in R — so u can activate v before other nodes in R do so. In addition, in a random permutation $\boldsymbol{\pi} \sim \Pi$ over V, the probability that $u \in R$ is ordered first in R is exactly $1/|R|$. This explains the contribution of $\mathbb{I}\{u \in \boldsymbol{R}\}/|\boldsymbol{R}|$ in Lemma 1, which is also precisely what the updates in lines 9 and 22 of Algorithm 1 do. The above two observations together establish Lemma 1, which is the basis for the unbiased estimator of u's Shapley centrality. Then, by a careful probabilistic analysis, we can bound the number of random RR sets needed to achieve approximation accuracy stated in Theorem 3 and establish the scalability for Algorithm ASV-RR.

Table 1: Datasets used in the experiments.

Dataset	# Nodes	# Edges	Weight Setting
Data mining (DM)	679	1687	WC, PR, LN
Flixster (FX)	29,357	212,614	LN
LiveJournal (LJ)	4,847,571	68,993,773	WC

4.2 Algorithm for SNI Centrality

Algorithm 1 relies on the key fact given in Lemma 1 about the Shapley centrality: $\psi_u^{Shapley} = n \cdot \mathbb{E}_{\boldsymbol{R}}[\mathbb{I}\{u \in \boldsymbol{R}\}/|\boldsymbol{R}|]$. A similar fact holds for the SNI centrality: $\psi_u^{SNI} = \sigma(\{u\}) = n \cdot \mathbb{E}_{\boldsymbol{R}}[\mathbb{I}\{u \in \boldsymbol{R}\}]$ [10, 32, 31]. Therefore, it is not difficult to verify that we only need to replace $\boldsymbol{est_u} = \boldsymbol{est_u} + 1/|\boldsymbol{R}|$ in lines 9 and 22 with $\boldsymbol{est_u} = \boldsymbol{est_u} + 1$ to obtain an approximation algorithm for SNI centrality. Let ASNI-RR denote the algorithm adapted from ASV-RR with the above change, and let ψ_v below denote SNI centrality ψ_v^{SNI} and $\psi^{(k)}$ denote the k-th largest SNI value.

THEOREM 4. *For any $\epsilon > 0$, $\ell > 0$, and $k \in \{1, 2, \ldots, n\}$, Algorithm ASNI-RR returns an estimated SNI centrality $\hat{\boldsymbol{\psi}}_v$ that satisfies (a) unbiasedness: $\mathbb{E}[\hat{\boldsymbol{\psi}}_v] = \psi_v, \forall v \in V$; and (b) robustness: with probability at least $1 - \frac{1}{n^\ell}$:*

$$\begin{cases} |\hat{\boldsymbol{\psi}}_v - \psi_v| \le \varepsilon\psi_v & \forall v \in V \text{ with } \psi_v > \psi^{(k)}, \\ |\hat{\boldsymbol{\psi}}_v - \psi_v| \le \varepsilon\psi^{(k)} & \forall v \in V \text{ with } \psi_v \le \psi^{(k)}. \end{cases} \quad (5)$$

Under Assumption 1 and the condition $\ell \ge (\log_2 k - \log_2 \log_2 n)/\log_2 n$, the expected running time of ASNI-RR is $O(\ell(m+n)\log n \cdot \mathbb{E}[\sigma(\tilde{\boldsymbol{v}})]/(\psi^{(k)}\varepsilon^2))$, where $\mathbb{E}[\sigma(\tilde{\boldsymbol{v}})]$ is the same as defined in Theorem 1.

Together with Algorithm ASV-RR and Theorem 3, we see that although Shapley and SNI centrality are quite different conceptually, surprisingly they share the same RR-set based scalable computation structure. Comparing Theorem 4 with Theorem 3, we can see that computing SNI centrality should be faster for small k since the k-th largest SNI value is usually larger than the k-th largest Shapley value.

5. EXPERIMENTS

We conduct experiments on a number of real-world social networks to compare their Shapley and SNI centrality, and test the efficiency of our algorithms ASV-RR and ASNI-RR.

5.1 Experiment Setup

The network datasets we used are summarized in Table 1.

The first dataset is a relatively small one used as a case study. It is a collaboration network in the field of Data Mining (DM), extracted from the ArnetMiner archive (arnetminer.org) [30]: each node is an author and two authors are connected if they have coauthored a paper. We use two large networks to demonstrate the effectiveness of the Shapley and SNI centrality and the scalability of our algorithms. Flixster (FX) [4] is a directed network extracted from movie rating site flixster.com. The nodes are users and a directed edge from u to v means that v has rated some movie(s) that u rated earlier. We use influence probabilities on topic 1 in their provided data as an example. Finally, LiveJournal (LJ) is the largest network we tested with. It is a directed network of bloggers, obtained from Stanford's SNAP project [1], and it was also used in [32, 31].

We use the independent cascade (IC) model in our experiments. The schemes for generating influence-probability profiles are also shown in Table 1, where WC, PR, and LN stand for *weighted cascade*, *PageRank-based*, and *learned from real data*, respectively. WC is a scheme of [18], which assigns $p_{u,v} = 1/d_v$ to edge $(u, v) \in E$, where d_v is the in-degree of node v. PR uses the nodes' PageRanks [11] instead of in-degrees: We first compute the PageRank score $r(v)$ for every node $v \in V$ in the unweighted network, using 0.15 as the restart parameter. Then, for each original edge $(u, v) \in E$, PR assigns an edge probability of $r(u)/(r(u) + r(v)) \cdot n/(2m^U)$, where m^U is the number of undirected edges in the graph. LN applies to DM and FX datasets, where we obtain learned influence probability profiles from the authors of the original studies. For the DM dataset, the influence probabilities on edges are learned by the topic affinity algorithm TAP proposed in [30]; for FX, the influence probabilities are learned using maximum likelihood from the action trace data of user rating events.

We implement all algorithms in Visual C++, compiled in Visual Studio 2013, and run our tests on a server computer with 2.4GHz Intel(R) Xeon(R) E5530 CPU, 2 processors (16 cores), 48G memory, and Windows Server 2008 R2 (64 bits).

An additional dataset DBLP with WC and PR settings is also tested and results are included in [12].

5.2 Experiment Results

Case Study on DM. We set $\varepsilon = 0.01$, $\ell = 1$, and $k = 50$ for both ASV-RR and and ASNI-RR algorithms. For the three influence profiles: WC, PR, and LN, Table 2 lists the top 10 nodes in both Shapley and SNI ranking together with their numerical values. The names appeared in all ranking results are well-known data mining researchers in the field,but the ranking details have some difference.

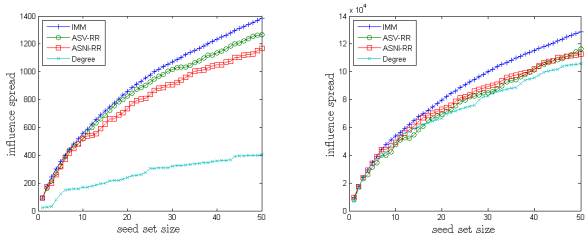
We compare the Shapley ranking versus SNI ranking under the same probability profiles. In general, the two top-10 ranking results align quite well with each other, showing that in these influence instances, high individual influence usually translates into high marginal influence. Some noticeable exception also exists. For example, Christos Faloutsos is ranked No.3 in the DM-PR Shapley centrality, but he is not in Top-10 based on DM-PR individual influence ranking. Conceptually, this would mean that, in the DM-PR model, Professor Faloutsos has better Shapley ranking because he has more unique and marginal impact comparing to his individual influence.

We next compare Shapley and SNI centrality with the structure-based degree centrality. The results show that the Shapley and SNI rankings in DM-WC and DM-PR are similar to the degree centrality ranking, which is reasonable because DM-WC and DM-PR are all heavily derived from node degrees. However, DM-LN differs from degree ranking a lot, since it is derived from topic modeling, not node degrees. This implies that when the influence model parameters are learned from real-world data, it may contain further information such that its influence-based Shapley or SNI ranking may differ from structure-based ranking significantly.

Results on Large Networks. We set $\varepsilon = 0.5$, $\ell = 1$, and $k = 50$ for this test, where $\varepsilon = 0.5$ is obtained with an omitted tuning step [12]. Due to the lack of user profiles, we assess the effectiveness of Shapley and SNI centralities through lens of influence maximization (IM), by comparing the IM performance of their top-ranked nodes with nodes selected by the IMM algorithm [31]. A simple heuristic Degree based on degree centrality is also compared as a baseline.

Table 2: Top 10 authors from DM dataset, ranked by Shapley, SNI, and degree centrality.

DM-WC				DM-PR				DM-LN					
Shapley		SNI		Shapley		SNI		Shapley		SNI		Degree	
Philip S. Yu	5.43	Philip S. Yu	40.59	Philip S. Yu	3.89	Philip S. Yu	61.14	Jiawei Han	23.27	Jiawei Han	51.01	Philip S. Yu	63
Jiawei Han	4.13	Jiawei Han	28.42	Jiawei Han	2.96	Jiawei Han	47.09	Qiang Yang	13.69	Qiang Yang	30.10	Jiawei Han	42
Wei Wang	3.96	Christos Faloutsos	23.89	Christos Faloutsos	2.64	Qiang Yang	40.98	Christos Faloutsos	10.92	Christos Faloutsos	22.89	Qiang Yang	34
Christos Faloutsos	3.83	Wei Wang	23.49	Heikki Mannila	2.61	Wei Wang	38.89	Heikki Mannila	10.40	Heikki Mannila	21.45	Christos Faloutsos	33
Heikki Mannila	3.48	Heikki Mannila	23.44	Wei Wang	2.50	Jian Pei	37.90	Vipin Kumar	7.99	Vipin Kumar	16.12	Heikki Mannila	33
Jian Pei	2.94	Qiang Yang	22.51	Qiang Yang	2.36	Vipin Kumar	36.81	C. Lee Giles	7.19	C. Lee Giles	14.54	Vipin Kumar	32
Qiang Yang	2.89	Jian Pei	21.63	Vipin Kumar	2.30	Bing Liu	35.86	Saso Dzeroski	7.16	Saso Dzeroski	14.50	Wei Wang	32
Vipin Kumar	2.85	Vipin Kumar	21.25	Jian Pei	2.25	Jeffrey Xu Yu	34.06	Graham J. Williams	6.71	Myra Spiliopoulou	13.39	Jian Pei	31
Bing Liu	2.84	Bing Liu	20.26	Bing Liu	2.15	Ke Wang	31.91	Eamonn J. Keogh	6.52	Eamonn J. Keogh	13.18	Bing Liu	28
C. Lee Giles	2.82	Ke Wang	17.89	Hiroshi Motoda	1.96	Hongjun Lu	30.16	Myra Spiliopoulou	6.43	Graham J. Williams	13.17	Ke Wang	26



(a) *Flixster-LN* (b) *LiveJournal-WC*

Figure 2: Influence maximization test.

Table 3: Running time (in seconds).

	ASV-RR	ASNI-RR	IMM
FX-LN	24.83	1.36	0.62
LJ-WC	8295.57	267.50	54.88

Figure 2 shows the influence spread results and Table 3 shows the running time. We see that both Shapley and SNI have similar IM performance and is reasonable close to the specialized IMM algorithm, but Shapley is noticeably better (average 8.3% improvement) than SNI in Flixster-LN test. This is perhaps due to that Shapley centrality accounts for more marginal influence, which is closer to what is needed for influence maximization. In FX-LN, both ASNI-RR and ASV-RR performs significantly better than Degree, again indicating that influence learned from the real-world data may contain significantly more information than the graph structure, in which case degree centrality is not a good index for node importance. Both ASNI-RR and ASV-RR can scale to the large LiveJournal group with $69M$ edges, and ASNI-RR scales better as predicted by Theorems 3 and 4.

6. CONCLUSION AND FUTURE WORK

Through an integrated mathematical, algorithmic, and empirical study of Shapley and SNI centralities in the context of network influence, we have shown that (a) both enjoy concise axiomatic characterizations, which precisely capture their similarity and differences; (b) both centrality measures can be efficiently approximated with guarantees under the same algorithmic structure, for a large class of influence models; and (c) Shapley centrality focuses on nodes' marginal influence and their irreplaceability in group influence settings, while SNI centrality focuses on individual influence in isolation, and is not suitable in assessing nodes' ability in group influence setting, such as threshold-based models.

There are several directions to extend this work and further explore the interplay between social influence and network centrality. One important direction is to formulate centrality measures that combine the advantages of Shapley and SNI centralities, by viewing Shapley and SNI centralities as two extremes in a centrality spectrum, one focusing on individual influence while the other focusing on marginal influence in groups of all sizes. Then, would there be some intermediate centrality measure that provides a better balance? Another direction is to incorporate other classical centralities into influence-based centralities. For example, SNI centrality may be viewed as a generalized version of degree centrality, because when we restrict the influence model to deterministic activation of only immediate neighbors, SNI centrality essentially becomes degree centrality. What about the general forms of closeness, betweenness, PageRank in the influence model? Algorithmically, efficient algorithms for other influence models such as general threshold models [18] is also interesting. In summary, this paper lays a foundation for the further development of the axiomatic and algorithmic theory for influence-based network centralities, which we hope will provide us with deeper insights into network structures and influence dynamics.

Acknowledgment

We thank Tian Lin for sharing his IMM implementation code and helping on data preparation. We thank David Kempe and four anonymous reviewers for their valuable feedback. Wei Chen is partially supported by the National Natural Science Foundation of China (Grant No. 61433014). Shang-Hua Teng is supported in part by a Simons Investigator Award from the Simons Foundation and by NSF grant CCF-1111270.

References

[1] Stanford network analysis project. https://snap.stanford.edu/data/.

[2] A. Altman and M. Tennenholtz. Ranking systems: The pagerank axioms. In *ACM*, EC '05, pages 1–8, 2005.

[3] K. J. Arrow. *Social Choice and Individual Values*. Wiley, New York, 2nd edition, 1963.

[4] N. Barbieri, F. Bonchi, and G. Manco. Topic-aware social influence propagation models. In *ICDM*, 2012.

[5] P. Boldi and S. Vigna. Axioms for centrality. *Internet Mathematics*, 10:222–262, 2014.

[6] P. Bonacich. Factoring and weighting approaches to status scores and clique identification,. *Journal of Mathematical Sociology*, 2:113–120, 1972.

[7] P. Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182, 1987.

[8] S. P. Borgatti. Centrality and network flow. *Social Networks*, 27(1):55–71, 2005.

[9] S. P. Borgatti. Identifying sets of key players in a social network. *Computational and Mathematical Organizational Theory*, 12:21–34, 2006.

[10] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier. Maximizing social influence in nearly optimal time. In *ACM-SIAM*, SODA '14, pages 946–957, 2014.

[11] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.

[12] W. Chen and S.-H. Teng. Interplay between social influence and network centrality: A comparative study on shapley centrality and single-node-influence centrality. *ArXiv e-prints*, (1602.03780), Feb. 2016.

[13] W. Chen, Y. Yuan, and L. Zhang. Scalable influence maximization in social networks under the linear threshold model. In *IEEE*, ICDM '10, pages 88–97, 2010.

[14] P. Domingos and M. Richardson. Mining the network value of customers. In *ACM*, KDD '01, pages 57–66, 2001.

[15] R. Ghosh and K. Lerman. Rethinking centrality: The role of dynamical processes in social network analysis. *Discrete and Continuous Dynamical Systems Series B*, pages 1355–1372, 2014.

[16] R. Ghosh, S.-H. Teng, K. Lerman, and X. Yan. The interplay between dynamics and networks: centrality, communities, and cheeger inequality. In *ACM*, KDD '14, pages 1406–1415, 2014.

[17] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, March 1953.

[18] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[19] T. P. Michalak, K. V. Aadithya, P. L. Szczepanski, B. Ravindran, and N. R. Jennings. Efficient computation of the shapley value for game-theoretic network centrality. *J. Artif. Int. Res.*, 46(1):607–650, Jan. 2013.

[20] R. B. Myerson. *Game Theory : Analysis of Conflict*. Harvard University Press, 1997.

[21] J. Nash. The bargaining problem. *Econometrica*, 18(2):155–162, April 1950.

[22] M. Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010.

[23] U. Nieminen. On the centrality in a directed graph. *Social Science Research*, 2(4):371–378, 1973.

[24] I. Palacios-Huerta and O. Volij. The measurement of intellectual influence. *Econometrica*, 72:963–977, 2004.

[25] M. Piraveenan, M. Prokopenko, and L. Hossain. Percolation centrality: Quantifying graph-theoretic impact of nodes during percolation in networks. *PLoS ONE*, 8(1), 2013.

[26] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *ACM*, KDD '02, pages 61–70, 2002.

[27] G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, 1966.

[28] D. Schoch and U. Brandes. Re-conceptualizing centrality in social networks. *European Journal of Applied Mathematics*, 27:971–985, 2016.

[29] L. S. Shapley. A value for n-person games. In H. Kuhn and A. Tucker, editors, *Contributions to the Theory of Games, Volume II*, pages 307–317. Princeton University Press, 1953.

[30] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *KDD*, 2009.

[31] Y. Tang, Y. Shi, and X. Xiao. Influence maximization in near-linear time: a martingale approach. In *SIGMOD*, pages 1539–1554, 2015.

[32] Y. Tang, X. Xiao, and Y. Shi. Influence maximization: near-optimal time complexity meets practical efficiency. In *SIGMOD*, pages 75–86, 2014.

[33] S.-H. Teng. Scalable algorithms for data and network analysis. *Foundations and Trends in Theoretical Computer Science*, 12(1-2):1–261, 2016.

[34] S.-H. Teng. Network essence: Pagerank completion and centrality-conforming markov chains. In J. N. Martin Loebl and R. Thomas, editors, *A Journey through Discrete Mathematics. A Tribute to Jiří Matoušek*. Springer Berlin / Heidelberg, 2017.

[35] C. Wang, W. Chen, and Y. Wang. Scalable influence maximization for independent cascade model in large-scale social networks. *DMKD*, 25(3):545–576, 2012.