Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

The Game-Theoretic Interaction Index on Social Networks with Applications to Link Prediction and Community Detection

Piotr L. Szczepański[1], **Aleksy Barcz**[1], **Tomasz P. Michalak**[2], **Talal Rahwan**[3]
[1]Warsaw University of Technology, 00-661 Warsaw, Poland
[2]University of Oxford, OX1 3QD, United Kingdom, University of Warsaw, 00-927 Warsaw, Poland
[3]Masdar Institute, Abu Dhabi 54224, UAE

Abstract

Measuring similarity between nodes has been an issue of extensive research in the social network analysis literature. In this paper, we construct a new measure of similarity between nodes based on the game-theoretic interaction index (Grabisch and Roubens, 1997). Despite the fact that, in general, this index is computationally challenging, we show that in our network application it can be computed in polynomial time. We test our measure on two important problems, namely link prediction and community detection, given several real-life networks. We show that, for the majority of those networks, our measure outperforms other local similarity measures from the literature.

1 Introduction

Vertex similarity is a fundamental network concept extensively studied in the literature [Blondel *et al.*, 2004; Jeh and Widom, 2002]. An effective measure of node-similarity allows for solving various real-life problems: establish trust-base recommendation [Hang and Singh, 2010], detect similar genes [Bass *et al.*, 2013], or recover damaged network [Chen *et al.*, 2012]. Two particularly interesting problems that are highly relevant to social networks are: (1) *link prediction* [Leicht *et al.*, 2006] and (2) *community detection* [Fortunato, 2010]. Specifically, the former problem involves estimating the likelihood of having a link between a given pair of nodes, based on the observed links and the attributes of nodes. The latter problem involves identifying groups of nodes such that there are relatively many links among nodes of the same group, and relatively few links among nodes belonging to different groups. Both problems are relevant to various fields in which graph representations are common, including sociology [Girvan and Newman, 2002] and computer science [Misra *et al.*, 2012].

There are two general classes of similarity measures: (1) *local*, and (2) *global*. The first class focuses on a local neighbourhood of the nodes in question (i.e., the nodes between which we are determining whether a link exists). Conversely, the second class considers the entire network. While global measures generally yield better qualitative results, they are more computationally involved, which limits their applicability to smaller networks. In contrast, local measures are scalable—they can be used even for very large networks. Thus, in this paper, we focus on local similarity measures.

The majority of local similarity measures are built upon the following concept: the more common neighbours two nodes share, the more similar those nodes are. This reasoning can be also extended to the entire neighbourhood of both nodes, or, in other words, their spheres of influence. Thus, similarity is proportional to the overlap of spheres of influence.

In this paper, we extend the above notion of similarity, building upon the game-theoretic *interaction index*—a concept developed in coalitional game theory and fuzzy systems to measure the interaction between a pair of elements embedded in a larger set of elements [Owen, 1972; Murofushi and Soneda, 1993; Grabisch and Roubens, 1999].

The basic form of the interaction index is built upon two well-known solution concepts from game theory—the Shapley value [Shapley, 1953b] and the Banzhaf index [Banzhaf, 1965]. It is also generalized to a wider class of solution concepts called *semivalues* [Grabisch and Roubens, 2000].

Our methodology may be summarised as follows. We first construct a coalitional game in which players are nodes in the network and the payoff of a coalition is a function of its sphere of influence parametrized by k, where k is the degree of influence. Given this influence game, we use the above interaction index to measure the similarity between the spheres of influence of any two nodes, given all the possible coalitions that those two nodes belong to.

Interestingly, despite the interaction index being hard to compute in general, we show that, in our influence game, it is computable in polynomial time. Specifically, the algorithm we propose for computing semivalue interaction indices runs in $O(|V|^2)$ time, after precomputations requiring $O(|V||E|+|V|^2)$. Better still, this polynomial-time result can be further improved upon if we reduce the scope of interaction in the network (defined by parameter k), or if we focus on the Shapley value-based interaction index. In the latter case computing similarities between two nodes requires $O(|V|)$ time, and this complexity is the same as for the fastest local similarity measures in literature. Overall, in the worst case scenario our link prediction and community detection algorithms based on the Shapley value-based interaction index run in $O(|V|^3)$ time.

We use our algorithm to study a number of real-life networks, and show that the interaction index-based similarity measure outperforms the other local measures in solving the link prediction and community detection problems.

2 Preliminaries

Let $N = \{1, 2, \ldots, n\}$ be the set of players. A *characteristic function*, ν, assigns a real *value* (or *payoff*) to every coalition of players: $\nu : 2^N \to \mathbb{R}$, with $\nu(\emptyset) = 0$. A tuple (N, ν) is called coalitional game.

Semivalues: One of the fundamental problems in cooperative game theory is how to evaluate the importance or contribution of players in a coalitional game. *Semivalues* represent an important class of solutions to this problem [Dubey *et al.*, 1981]. To define semivalues, let us first denote by $\text{MC}(C, i)$ the marginal contribution of the player i to coalition C, i.e., $\text{MC}(C, i) = \nu(C \cup \{i\}) - \nu(C)$. Furthermore, let $\alpha : \{0, 1, \ldots, |N| - 1\} \to [0, 1]$ be a discrete probability distribution. Intuitively, $\alpha(k)$ is the probability that a coalition of size k is drawn from the set of all possible coalitions. Now, given a function α, the semivalue $\phi_v(\nu)$ of a player i in the cooperative game ν is:

$$\phi_i(\nu) = \sum_{0 \le l < |N|} \alpha(l) \mathbb{E}[\text{MC}(C^l, i)], \tag{1}$$

where C^l is the random variable that corresponds to a coalition being drawn with uniform probability from the set of all coalitions of size l in the set of players $N \setminus \{i\}$, while $\mathbb{E}[\cdot]$ is the expected value operator. For instance, some prominent semivalues include the *Shapley value* [Shapley, 1953a] and the Banzhaf index [Banzhaf, 1965]—two solution concepts widely studied in cooperative game theory due to their various desirable properties [Maschler *et al.*, 2013].

Interaction index: To understand the notion of the interaction index, we have to first introduce the definition of synergy between two players in cooperative games. It is the additional value (either positive or negative) that these players achieve by cooperating. Formally: $\text{S}(i, j) = \nu(\{i, j\}) - \nu(\{i\}) - \nu(\{j\})$. Similarly, the synergy between i and j with respect to a coalition C is:

$$\text{S}(C, i, j) = \text{MC}(C, \{i, j\}) - \text{MC}(C, i) - \text{MC}(C, j).$$

The notion of *interaction* among entities, or "*players*", builds upon the above notion of synergy. Apart from game theory, it has been studied in various other fields, such as fuzzy systems, multi-criteria decision making, aggregation function theory, statistics and data analysis [Marichal and Mathonet, 2008]. Informally, the value of interaction among a group of players is the average synergy of this group over all coalitions they could potentially belong to. In this paper we focus on the semivalue interaction index between two players, which is defined as follows:

Definition 1 *The semivalue interaction index in (N, ν) is:*

$$I_{ij}(N, \nu) = \sum_{0 \le l \le |N| - 2} \beta(k) \mathbb{E}[S(C^l, i, j)], \tag{2}$$

where $\beta : \{0, 1, \ldots, |N| - 2\} \to [0, 1]$ is a discrete probability distribution and $\mathbb{E}[S(C^l, i, j)] = \sum_{C \in \{C \subseteq N \setminus \{i,j\} \,:\, |C| = l\}} \frac{S(C,i,j)}{\binom{|V|-2}{l}}$.

The interpretation of these index is as follows: **a)** if $\text{I}_{ij} < 0$ then i and j have an overall negative influence on each other, **b)** if $\text{I}_{ij} > 0$ then i and j have an overall positive influence on each other, **c)** if $\text{I}_{ij} = 0$ then i and j do not influence each other, or their influences cancel out. Generally, if I_{ij} is very high, or low, the players interact between themselves intensively. The three semivalue interaction indices widely studied in literature are: Shapley interaction index [Grabisch, 1997] ($\beta(l) = 1/(|V| - 1)$), Banzhaf interaction index [Roubens, 1996] ($\beta(l) = \binom{|V|-2}{l} / 2^{|V|-2}$) and chaining interaction index [Marichal and Roubens, 1999] ($\beta(l) = \frac{2(l+1)}{(n-1)(n-2)}$).

3 Our Model

In this section we describe our approach to define new similarity function. First, we introduce the influence games that underpin the interaction index. Next, we define our key concept—the semivalue k-steps interaction index defined on networks. Finally, we translate desirable properties of this index from game-theory to the network context.

Influence games: Michalak et al. (2013b) studied a number of coalition games defined on networks, in which the characteristic function evaluates each subset of nodes proportionally to the size of the sphere of influence that this subset has in this network. If we denote the sphere of influence of the set $C \subseteq V$ as $SF(C) \subseteq V$ then:

Definition 2 *The influence game of a graph $G = (V, E)$ is a tuple (G, ν_G), where $\nu_G : 2^V \to \mathbb{R}$ is an influence function such than for each $C \subseteq V$ we have $\nu_G(C) = SF(C)$.*

In this paper we focus on the k-steps influence function.

Definition 3 *The k-steps influence function is:*

$$\nu_k(S) = |\{v \in V \mid d(S, v) < k\} \setminus S|, \tag{3}$$

where $d(v, u)$ is the distance between two nodes and $d(S, v) = \min_{u \in S} d(u, v)$.

The interaction index on networks: In further sections we use the following interaction index in order to measure similarity between nodes in networks:

Definition 4 *The semivalue k-steps interaction index is a tuple (G, ν_k), where $G = (V, E)$, ν_k is the k-step influence function, and the interaction between any two nodes $u, v \in V$ is given by $I_{uv}(V, \nu_k)$.*

The properties: Now, we translate the game-theoretic axiomatization of the semivalue interaction index to the network context [Grabisch and Roubens, 1999].

Property 1 *Interaction between nodes is symmetric. That is, $\forall_{v,u \in V} I_{uv} = I_{vu}$.*

Property 2 *If a node always contributes its singleton value to any community, then its interaction is always zero:*[1] $\forall_{C \subseteq V_{\setminus \{v\}}} \text{MC}(C, v) = \nu_G(\{v\}) \Longrightarrow \forall_{u \in V_{\setminus \{v\}}} I_{vu} = 0$.

[1] Intuitively, if a node always contributes the same, its sphere does not overlap with that of any coalition, and so it is considered to have zero interaction with others.

Property 3 *If two influence functions are combined into one* $\nu_G = \nu'_G + \nu''_G$ *then* $I_{uv}(\nu_G) = I_{uv}(\nu'_G) + I_{uv}(\nu''_G)$.

Property 4 *If two nodes contribute the same value to all communities then they have the same interaction:* $\forall_{C \subseteq V_{\setminus\{v,u\}}} \text{MC}(C, v) = \text{MC}(C, u) \Longrightarrow \forall_{w \in V_{\setminus\{v,u\}}} I_{vw} = I_{uw}$.

4 Computing interaction index on networks

In this section, we tackle the problem of computing the k-steps semivalue interaction index on large networks. Despite the exponential space of all subsets of nodes in the network, we present a combinatorial analysis that results in an algorithm capable of measuring the interaction between any two nodes in polynomial time. Our analysis is motivated by the work done by Szczepański et al (2015), who defined the class of polynomial-time computable game-theoretic network centralities. Here, we focus on equation (2). More specifically, for each pair of nodes $v, u \in V$ we will show how to compute $\mathbb{E}[\text{S}(C^l, v, u)]$—the expected value of their synergy with respect to the random set C^l. Our main observation is that this computation comes down to computing the expected marginal contribution of v, and of u, to C^l. Formally:

$$\mathbb{E}[\text{S}(C^l, u, v)] = \mathbb{E}[\text{MC}(C^l, \{u, v\})] - \mathbb{E}[\text{MC}(C^l, u)] - \mathbb{E}[\text{MC}(C^l, v)] \quad (4)$$

Before presenting our main theorem, we need additional notation. For every node $v \in V$, let $N_k(v)$ denote the *set* of nodes reachable from v with at most k steps, and let $deg_k(v)$ denote the *number* of such nodes. Formally, $N_k(v) = \{u \in V \mid d(v, u) \leq k \wedge v \neq u\}$ and $deg_k(v) = |N_k(v)|$. We extend this notation to sets of nodes. That is, $N_k(C) = \bigcup_{v \in C} N_k(v) \setminus C$ and $deg_k(C) = |N_k(C)|$. Now, we are ready to introduce the following theorem:

Theorem 1 *Given a graph, G, and a pair of nodes* $u, v \in V$*, the* k*-steps semivalue interaction index between* u *and* v *can be computed in time polynomial in* $|V|$.

Proof: Let us focus on one of the two nodes, say v, and analyse its marginal contribution to $C^l \subseteq V \setminus \{v, u\}$. In this analysis, we will first focus on the cases where the marginal contribution is positive, and then move our attention to the cases where it is negative. Regarding the case of positive marginal contribution, when v joins C^l, such a contribution is only made with the help of another node $n \in N_k(v)$ where n is not in C^l nor is it accessible within k steps from C^l. Such a contribution is indeed positive because the node n is not under the influence of coalition C^l but under the influence of coalition $C^l \cup \{v\}$. Moving on to the case of negative marginal contributions, such a case only happens when v is accessible within k steps from C^l. This is because the coalition C^l has influence on the node v while the coalition $C^l \cup \{v\}$ has no influence. In order to formalize the above observations, we introduce two Bernoulli random variables. The first one, denoted by $B^+_{l,v,n}$, indicates whether the node v makes a positive marginal contribution through node n to the random set C^l. Formally, we have:

$$\mathbb{E}[B^+_{l,v,n}] = P[(N_k(n) \cup \{n\}) \cap C^l = \emptyset], \quad (5)$$

where $P[\cdot]$ denotes probability, and $\mathbb{E}[\cdot]$ denotes expected value. The second one, denoted by $B^-_{l,v}$, indicates whether the node v makes a negative marginal contribution through itself to the random set C^l. More formally:

$$\mathbb{E}[B^-_{l,v}] = P[N_k(v) \cap C^l \neq \emptyset]. \quad (6)$$

Now, we will use a combinatorial argument to compute the probabilities in equations (5) and (6). For the sake of clarity we will assume that $\binom{a}{b} = 0$ for $a < b$ and that for any a we have $\frac{a}{0} = 0$. Recall that there are exactly $\binom{|V|-2}{l}$ possible sets C^l. With this in mind, we will start by showing how the probability in (5) is computed. Here, we will focus on the following two complementary cases:

- Case 1: $n \in (N_k(v) \setminus \{u\})$ such that $n \neq u$ and $u \notin N_k(n)$. In this case, the probability in (5) can be computed as follows:

$$P[(N_k(n) \cup \{n\}) \cap C^l = \emptyset] = \frac{\binom{|V|-2-deg_k(n)}{l}}{\binom{|V|-2}{l}}.$$

  The nominator of the above fraction indicates that, in order to satisfy the condition $(N_k(n) \cup \{n\}) \cap C^l = \emptyset$, we can choose any of the nodes in $V \setminus \{u, v\}$ except those that are in $n \cup (N_k(n) \setminus \{v\})$.

- Case 2: $n \in (N_k(v) \setminus \{u\})$ such that $n = u$ or $u \in N_k(n)$. In this case, since we have $u \notin C^l$, the probability in (5) is computed differently as follows:

$$P[(N_k(n) \cup \{n\}) \cap C^l = \emptyset] = \frac{\binom{|V|-2-deg_k(n)+1}{l}}{\binom{|V|-2}{l}},$$

By combining the above two cases, for v and any $n \in N_k(v)$ we obtain:

$$\mathbb{E}[B^+_{l,v,n}] = \begin{cases} \frac{\binom{|V|-1-deg_k(n)}{l}}{\binom{|V|-2}{l}} & \text{if } n = u \text{ or } u \in N_k(n) \\ \frac{\binom{|V|-2-deg_k(n)}{l}}{\binom{|V|-2}{l}} & \text{otherwise.} \end{cases} \quad (7)$$

Having shown how to compute $\mathbb{E}[B^+_{l,v,n}]$, we now move to $\mathbb{E}[B^-_{l,v}]$. In particular, we show how to compute the probability in (6). To this end, observe that $P[N_k(v) \cap C^l \neq \emptyset] = 1 - P[N_k(v) \cap C^l = \emptyset])$. Based on this, instead of focusing on the probability of the event $N_k(v) \cap C^l \neq \emptyset$, we focus on the probability of the complementary event, i.e., $N_k(v) \cap C^l = \emptyset$, as this simplifies our analysis. Using the same combinatorial argument as before, we obtain the following for v:

$$\mathbb{E}[B^-_{l,v}] = \begin{cases} 1 - \frac{\binom{|V|-1-deg_k(v)}{l}}{\binom{|V|-2}{l}} & \text{if } u \in N_k(v) \\ 1 - \frac{\binom{|V|-2-deg_k(v)}{l}}{\binom{|V|-2}{l}} & \text{otherwise.} \end{cases} \quad (8)$$

The final formula for $\mathbb{E}[\text{MC}(C^l, v)]$ combines equations (7) and (8). That is,

$$\mathbb{E}[\text{MC}(C^l, v)] = \sum_{n \in N_k(v)} \Big(\mathbb{E}[B^+_{l,v,n}]\Big) - \mathbb{E}[B^-_{l,v}]. \quad (9)$$

So far, we have shown how to compute one of the three expressions in equation (4), namely $\mathbb{E}[\text{MC}(C^l, v)]$. Moving on the expression $\mathbb{E}[\text{MC}(C^l, u)]$ in (4), this can be computed in exactly the same manner. Finally, the third expression in (4), i.e., $\mathbb{E}[\text{MC}(C^l, \{u, v\})]$, is computed slightly differently. For each $n \in N_k(\{u, v\})$ we have:

$$\mathbb{E}[B^+_{l,\{v,u\},n}] = \begin{cases} \frac{\binom{|V|-1-deg_k(n)}{l}}{\binom{|V|-2}{l}} & \text{if } n \in (N_k(u) \cap N_k(v)) \\ \frac{\binom{|V|-2-deg_k(n)}{l}}{\binom{|V|-2}{l}} & \text{otherwise,} \end{cases} \tag{10}$$

and we also have:

$$\mathbb{E}[B^-_{l,\{v,u\}}] = \mathbb{E}[B^-_{l,u}] + \mathbb{E}[B^-_{l,v}] \tag{11}$$

Finally, we can combine all equations in order to compute the expected synergy:

$$\begin{aligned} \mathbb{E}[\text{S}(C^l, u, v)] &= \sum_{n \in N_k(\{v,u\})} \Big(\mathbb{E}[B^+_{l,\{v,u\},n}]\Big) - \mathbb{E}[B^-_{l,\{v,u\}}] \\ &- \sum_{n \in N_k(u)} \Big(\mathbb{E}[B^+_{l,u,n}]\Big) + \mathbb{E}[B^-_{l,u}] - \sum_{n \in N_k(v)} \Big(\mathbb{E}[B^+_{l,v,n}]\Big) + \mathbb{E}[B^-_{l,v}] \\ &= \sum_{\substack{n \in \left(N_k(v) \cap N_k(u)\right) \cup \\ \left(\{u\} \cap N_k(v)\right) \cup \left(\{v\} \cap N_k(u)\right)}} \Big(- \frac{\binom{|V|-1-deg_k(n)}{l}}{\binom{|V|-2}{l}} \Big) \end{aligned} \tag{12}$$

The above formula can be used to compute $\mathbb{E}[\text{S}(C^l, u, v)]$ in polynomial time. Therefore, the k-steps semivalue interaction index between v and u can be computed in time polynomial in $|V|$ using equation (2), which ends our proof. □

4.1 Algorithm

Algorithm 1 directly implements expression (2). The expected value operator $\mathbb{E}[\text{S}(C^l, u, v)]$ is computed using equation (12). It computes the k-steps semivalue interaction index for a given pair of nodes $u, v \in V$ in the graph G. This algorithm requires some precomputations. For each node $u \in V$ we need to calculate $N_k(u)$. We can store these values using $O(|V|^2)$ space and perform these precomputations in $O(|V|(|V| + |E|))$ time using breadth-first search. After this precomputation, the algorithm itself runs in $O(|V|^2)$ time. Our algorithm can be easily adapted to weighted graphs; the only difference is that the precomputations should be carried out using the Dijkstra algorithm (which takes $O(|V|(|E| + |V| \log |V|))$ time). The algorithm itself remains unchanged.

Algorithm 1: k-steps semivalue interaction index

Input: Graph $G = (V, E)$, nodes $v, u \in V$, functions β
Data: for $u \in V$: $N_k(u)$ - the set of k-Neighbours
Output: $\text{I}^k_{u,v}$ k-steps semivalue interaction index

$\text{I}^k_{u,v} \leftarrow 0$;
for $l \leftarrow 0$ **to** $|V| - 2$ **do**
  $\text{S}_{l,u,v} \leftarrow 0$;
  foreach $n \in \big(N_k(v) \cap N_k(u)\big)$ **do**
    $\text{S}_{l,u,v} \leftarrow \text{S}_{l,u,v} - \binom{|V|-1-deg_k(n)}{l}$;
  if $v \in N_k(u)$ **then**
    $\text{S}_{l,u,v} \leftarrow \text{S}_{l,u,v} - \binom{|V|-1-deg_k(u)}{l} - \binom{|V|-1-deg_k(v)}{l}$;
  $\text{I}^k_{u,v} \leftarrow \text{I}^k_{u,v} + \frac{\beta(l)}{\binom{|V|-2}{l}} \text{S}_{l,u,v}$;

Algorithm 2: k-steps Shapley interaction index

Input: Graph $G = (V, E)$, nodes $v, u \in V$
Data: for each node $u \in V$:
$N_k(u)$ - the set of k-Neighbours
Output: $\text{I}^k_{u,v}$ k-steps Shapley value interaction index

$\text{I}^{SV,k}_{u,v} \leftarrow 0$;
foreach $n \in \big(N_k(v) \cap N_k(u)\big)$ **do**
  $\text{I}^{SV,k}_{u,v} \leftarrow \text{I}^{SV,k}_{u,v} - \frac{1}{deg_k(n)-1}$;
if $v \in N_k(u)$ **then**
  $\text{I}^{SV,k}_{u,v} \leftarrow \text{I}^{SV,k}_{u,v} - \frac{1}{deg_k(u)-1} - \frac{1}{deg_k(v)-1}$;

We note that our algorithm can be optimized for any specific semivalue interaction index. For instance, Algorithm 2 is a version of our algorithm that computes the k-steps Shapley value interaction index in $O(|V|)$ time.

Our final observation is that Algorithms 1 and 2 always result in non-positive values of interaction between nodes. This is due to the fact, that our coalitional game defined on graph (G, ν_k) is weakly subadditive, that is in our game only negative synergies can occur. So, the more negative value $\text{I}^{SV,k}_{u,v}$ is, the more similar nodes u and v are.

5 The Performance of Interaction Index

In this section, we evaluate the effectiveness of our similarity measure based on the Shapley k-steps interaction index. More specifically, we compare it against the other measures from Table 1, which are all based on both the topology of the network (known as *structural similarities*) and local information (known as *local similarities*) [Lü and Zhou, 2011]. In all of these measures, a parameter k is used to define the influence of v, denoted by $N_k(v)$. This is simply the set of nodes that are k steps away from v in the network. That is to say, every node v is assumed to have k degrees of influence. In our experiments we test Shapley value-based interaction index as it is the most prominent semivalue and it has been already widely studied in the context of social networks [Michalak *et al.*, 2013b; Szczepański *et al.*, 2014].

5.1 Link Prediction

Having proposed a new algorithm for computing vertex similarity, in this subsection we show how to use it to solve the link prediction problem on various real-life networks.

Link prediction and similarity measure: We evaluate the effectiveness of each similarity measure using the following standard procedure for solving the link prediction problem: given a graph G with missing or not-yet-developed edges, the procedure invovles two steps: (i) compute the similarity of

Name	Measure
Common Neighbors **(CN)**:	$S^{CN}_{u,v} = N_k(u) \cap N_k(v)$
Salton Index **(SI)**:	$S^{SI}_{u,v} = \frac{N_k(u) \cap N_k(v)}{\sqrt{N_k(u) \times N_k(v)}}$
Jaccard Index **(JI)**	$S^{JI}_{u,v} = \frac{N_k(u) \cap N_k(v)}{N_k(u) \cup N_k(v)}$
Hub Promoted Index **(HPI)**	$S^{HPI}_{u,v} = \frac{N_k(u) \cap N_k(v)}{\min(N_k(u), N_k(v))}$
Leicht-Holme-Newman Index **(LHN)** :	$S^{LHN}_{u,v} = \frac{N_k(u) \cap N_k(v)}{N_k(u) \times N_k(v)}$
Adamic-Adar Index **(AA)** :	$S^{AA}_{u,v} = \sum_{n \in N_k(u) \cap N_k(v)} \frac{1}{\log N_k(n)}$
Resource Allocation **(RA)** :	$S^{RA}_{u,v} = \sum_{n \in N_k(u) \cap N_k(v)} \frac{1}{N_k(n)}$

Table 1: The seven measures used in our experiments.

each pair of nodes that have no edge between them, and (ii) connect the most similar pairs.

Network	SvII	CN	SI	JI	HPI	LHN	AA	RA
	PR AUC	PR AUC	PR AUC	PR AUC	PR AUC	PR AUC	PR AUC	PR AUC
Zachary	**278 758**	255 655	143 627	148 702	144 566	132 553	265 748	248 744
Dolphins	**187 793**	168 743	039 676	174 770	034 657	023 674	172 774	176 766
PolBooks	**275 829**	265 802	160 802	209 808	166 795	122 762	270 818	270 824
Football	388 913	388 890	438 **918**	442 916	429 906	**449** 912	380 906	385 904
Proteins	032 **630**	**041** 525	013 373	033 531	016 374	016 372	035 531	032 533
Emails	**123 817**	095 766	002 759	089 808	010 724	010 652	101 780	119 806
Powergrid	**042 661**	034 616	005 510	032 616	005 512	005 512	035 617	035 615

Table 2: The precision (PR) and arena under curve (AUC) of the effectiveness of different measures in link prediction on seven real-life networks. If PR equals 1000 than all missing edges were detected, and if AUC is 1000 than all original edges from G all rank higher than not existing connections.

Experiment: Our experiment was performed on seven popular real-life networks.[2] The basic idea is to remove some edges from the network under consideration, then test the effectiveness of each similarity measure in predicting where the missing edges should be. More specifically, the settings were as follow:

- We randomly remove 40%[3] of edges from the original graph $G = (V, E)$, thus obtaining the training graph $G^T = (V, E^T)$.
- For each node $v \in V$ we create 8×3 different rankings based on 8 similarity measures (the Shapley k-steps interaction index , and the seven measures from Table 1) and $k \in \{1, 2, 3\}$. Specifically, for each such configuration, we rank all nodes that are not directly connected to v. Then, we take the number of missing edges, i.e., $deg_G(v) - deg_{G^T}(v)$, from the top of the ranking, and compute the precision. The AUC [Lü and Zhou, 2011] is computed for the entire ranking.
- For each of the real-world networks in our experiments, the above process is repeated 30 times, and the average precision and AUC is computed.

Table 2 presents the results of the experiment. Here, the bold text highlights the best performance out of all measures. As can be seen, our new measure SvII outperforms the others on most networks.

Interestingly, although our measure is very similar to the Resource Allocation measure (RA) [Zhou *et al.*, 2009], ours

[2]All datasets are at *http://konect.uni-koblenz.de/networks/*.

[3]Lower percentages 20% and 10% were also tested resulted in higher precision and AUC and similar performance rankings.

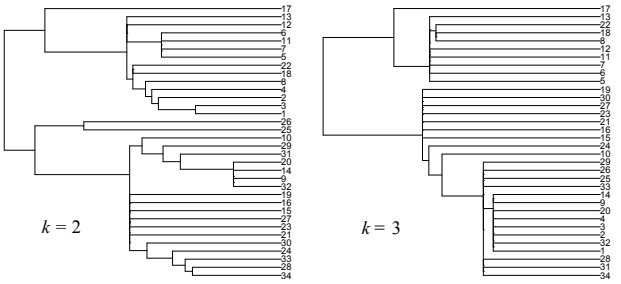

Figure 1: Dendrograms of the Zachary karate club network for $k = 2$ and $k = 3$, respectively.

dominates RA on all networks. Perhaps the reason behind this difference is the following: when measuring the similarity between v and u, the RA measure assumes that node v sends some resource to u through a common neighbor $n \in N_k(u) \cap N_k(v)$, who simply divides this resource equally among the nodes in $N_k(n)$, meaning that u recieves exactly $\frac{1}{deg_k(n)}$ of the resource. On the other hand, when using our measure, the fraction in formula (12) is $\frac{1}{deg(n)-1}$, not $\frac{1}{deg(n)}$, which may indicate that the node n does not send back any of the resource to the sender, v. Another difference compared to the RA measure is that, with our measure, if nodes u and v happen to be within each others‘ sphere of influence, i.e., $u \in N_k(v)$, then they can transmit resource directly to each other (see formula (12) and Algorithm 2).

5.2 Community detection algorithm

In this subsection we show how to use our interaction index to identify communities.

Hierarchical clustering: The effectiveness of each similarity measure is evaluated using the following standard procedure for detecting communities: starting with $|V|$ communities, each containing a single node, the algorithm proceeds by (i) merging the two most similar communities, and (ii) evaluating the *modularity* [Newman, 2006] of the resulting community structure. This procedure of merging communities and evaluating modularity is repeated until we end up with a single community containing all nodes. Finally, the community structure with the highest modularity is chosen. We will refer to this procedure as the bottom-up hierarchical clustering algorithm, or simply clustering algorithm, for community detection. Note that step (i) requires measuring the similarity between two *groups* of nodes, not two *individual* nodes as is the case with all aforementioend measures of similarity. To address this issue, we generalize each measure based on the notion of *complete* linkage (furthest-neighbour) [Jain *et al.*, 1999], which basically states that an individual should not join a community if his interaction with any of the community members is too small. We are looking for communities containing tightly connected nodes and this is exactly done by complete linkage: ”The complete-link algorithm produces tightly bound or compact clusters” [Jain *et al.*, 1999] Other popular linkage types, namely *single* linkage and *average* linkage, were also tested. However, the *complete* linkage proved to yield the best results for all the similarity measures.

The above community detection algorithm runs in $O(|V|^2 \log |V|)$ time, provided that the similarity between

Network	SvII	CN	SI	JI	HPI	LHN	AA	RA
	mod cov	mod cov	mod cov	mod cov	mod cov	mod cov	mod cov	mod cov
Zachary	**391** 628	313 846	301 808	301 808	313 846	310 666	313 846	313 846
Dolphins	447 679	**464** 811	381 730	379 723	370 660	445 698	421 736	442 780
Books	445 946	445 946	445 946	445 946	457 796	**481** 864	448 950	445 946
Football	**601** 690	**601** 690	**601** 690	**601** 690	**601** 690	**601** 690	**601** 690	586 674
Proteins	**741** 824	562 652	524 865	601 684	677 746	591 660	695 803	705 807
Emails	**397** 471	354 475	339 457	256 447	256 447	252 287	322 428	392 473
Powergrid	**864** 902	841 884	834 879	824 869	847 883	**864** 897	846 889	860 900

Table 3: Modularity and coverage for hierarchical community detection algorithm with different interaction measures.

Figure 2: The community structures with optimal modularity for the Political Books network ($k = 1, 2$, respectively).

each pair of nodes has been precomputed, which takes $O(|V|^3)$ time regardless of the measure used.

Next, we perform two experiments. In the first, we evaluate the sensitivity of the clustering algorithm to the parameter k when using our Shapley k-steps interaction index. In the second experiment, we evaluate the effectiveness of our measure, compared to those in Table 1, in terms of the modularity of the resulting community structure.

Experiment 1: the parameter k. Generally speaking, our measure, as well as any of the other measures from Table 1, can be controlled by the parameter k, which specifies the degree of influence in the network. This is particularly useful for community detection, as it allows for controlling the size of the resulting communities. In other words, by using any such measure, we obtain a *multi-resolution method* [Fortunato, 2010]. We analyze the outcome given different values of k and given our measure of similarity: if k is too small, then some potentially important interactions may not be detected. On the other hand, if k is too large, the interaction between weakly related nodes may negatively affect the clustering process.

Let us illustrate what happens when k is too large on the widely studied friendship network of the Zachary karate club [Zachary, 1977]. Figure 1 presents the hierarchical clustering of this network. We observe that for $k = 3$ the dendrogram has fewer levels compared to $k = 2$. This means that, for $k = 3$, in each step of the clustering process, the interaction computed between many pairs of clusters (possibly single nodes) is the same. Thus, for this network, $k = 3$ is too high, and results in a lower modularity compared to $k = 2$.

Next, let us consider what happens when k is too small. In Figure 2 we can see two community structures for the Political Books network, in which an edge between two books indicates that they were co-purchased by the same buyer on Amazon [Krebs, 2004]. For $k = 2$ we obtained a division into two communities: (roughly speaking) books supporting government in one community, and criticising it in the other. However, if we set $k = 1$ we obtain a division into 12 communities. This latter community structure turns out to be much worse in terms of modularity than the former one.

Generally speaking, it certainly helps the algorithm to try out all possible k and choose the one that yields the best community structure. However, this increases the computational complexity, and slows down the algorithm especially given large networks.

Experiment 2: performance. Now, the performance of the clustering algorithm given our similarity measure is compared against its performance given other measures from Table 1. Importantly, the ground-truth community structure is unknown for all networks considered in this experiment, with the only exception being the Zachary network. Therefore, following the standard practice in the literature, the quality of any given community structure will be measured using *modularity* [Newman, 2006].

The results are reported in Table 3. The table also reports the coverage [Brandes and Erlebach, 2005] of the selected community structure, which simply measures the ratio between intra-edges (those within a community) and inter-edges (those between communities). As can be seen, for most of the networks, the clustering algorithm using our similarity measure (SvII) gives the best results. Once again, ours dominates the Resource Allocation (RA) measure on all datasets. The second best measure is Leicht-Holme-Newman Index (LHN) [Leicht *et al.*, 2006]. It gives high similarity to nodes that have many common neighbors compared to the expected number of such neighbors.

6 Summary & Future Work

In this paper we proposed a new vertex similarity measure based on the game-theoretic notion of the interaction index. Despite the computational challenges posed by the combinatorial nature of this measure, we showed that, given our influence game, it is possible to compute this index in polynomial time. We tested our measure on two important applications: link prediction and community detection. For both applications, our measure outperformed other alternatives from literature.

While in this work we focused on one particular influence game, it would be interesting to study other, perhaps more involved but still polynomially-computable ones like those based on betweenness centrality [Szczepański *et al.*, 2012] or connectivity games [Michalak *et al.*, 2013a]. Furthermore, while we focused on standard games, it would be interesting to study interaction index defined for generalized coalitional games [Michalak *et al.*, 2014] but extended to graphs.

Acknowledgments

Piotr Szczepański was founded by the Polish National Science Centre based on the decision DEC-2013/09/N/ST6/04095. Tomasz Michalak was supported by the European Research Council under Advanced Grant 291528 ("RACE")

References

[Banzhaf, 1965] J. F. Banzhaf. Weighted Voting Doesn't Work: A Mathematical Analysis. *Rutgers Law Review*, 19:317–343, 1965.

[Bass *et al.*, 2013] J. I. Bass, A. Diallo, J. Nelson, J. M. Soto, C. L. Myers, and A. J. M. Walhout. Using networks to measure similarity between genes: association index selection. *Nat Meth*, 10(12):1169–1176, December 2013.

[Blondel *et al.*, 2004] V. D. Blondel, A. Gajardo, M. Heymans, P. Senellart, and P. Van Dooren. A measure of similarity between graph vertices: App. to synonym extraction and web searching. *SIAM Rev.*, 46(4):647–666, 2004.

[Brandes and Erlebach, 2005] U. Brandes and T. Erlebach. *Network Analysis: Methodological Foundations (Lecture Notes in Computer Science)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[Chen *et al.*, 2012] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles. Discovering missing links in networks using vertex similarity measures. In *Proc. of the 27th Annual ACM SAC*, SAC '12, pages 138–143, 2012.

[Dubey *et al.*, 1981] P. Dubey, A. Neyman, and R. J. Weber. Value Theory Without Efficiency. *Mathematics of Operations Research*, 6:122–128, 1981.

[Fortunato, 2010] S Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.

[Girvan and Newman, 2002] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002.

[Grabisch and Roubens, 1999] M. Grabisch and M. Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28:547–565, 1999.

[Grabisch and Roubens, 2000] M. Grabisch and M. Roubens. *Probabilistic interactions among players of a cooperative game*. 2000.

[Grabisch, 1997] M. Grabisch. k-order additive discrete fuzzy measures and their representation. *Fuzzy Sets and Systems*, 92(2):167 – 189, 1997.

[Hang and Singh, 2010] C. Hang and M. P. Singh. Trust-based recommendation based on graph similarity. In *in AAMAS W. on Trust in Agent Societies (Trust*, 2010.

[Jain *et al.*, 1999] A. K. Jain, M N. Murty, and P. J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[Jeh and Widom, 2002] G. Jeh and J. Widom. Simrank: A measure of structural-context similarity. In *Proc. of the Eighth ACM SIGKDD*, KDD '02, pages 538–543, 2002.

[Krebs, 2004] V. Krebs. Books about us politics, 2004.

[Leicht *et al.*, 2006] E. A. Leicht, Petter Holme, and M. E. J. Newman. Vertex similarity in networks. *Phys. Rev. E*, 73:026120, Feb 2006.

[Lü and Zhou, 2011] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A*, 390(6):11501170, 2011.

[Marichal and Mathonet, 2008] J.-L. Marichal and P. Mathonet. Approximations of lovsz extensions and their induced interaction index. *DAM*, 156(1):11–24, 2008.

[Marichal and Roubens, 1999] J. L. Marichal and M. Roubens. The chaining interaction index among players in cooperative games. In *Adv. in Decision Analysis*, volume 4 of *Math.Mod.-TA*, pages 69–85. 1999.

[Maschler *et al.*, 2013] M. Maschler, S. Zamir, and E. Solan. *Game Theory*. Cambridge University Press, 2013.

[Michalak *et al.*, 2013a] T. P. Michalak, T. Rahwan, P. L. Szczepanski, O. Skibski, R. Narayanam, M. J. Wooldridge, and N. R. Jennings. Computational analysis of connectivity games with applications to the investigation of terrorist networks. *IJCAI*, 2013.

[Michalak *et al.*, 2013b] Tomasz P. Michalak, Karthik V. Aadithya, Piotr L. Szczepanski, Balaraman Ravindran, and Nicholas R. Jennings. Efficient computation of the shapley value for game-theoretic network centrality. *J. Artif. Intell. Res. (JAIR)*, 46:607–650, 2013.

[Michalak *et al.*, 2014] T. P. Michalak, P. L. Szczepański, T. Rahwan, A. Chrobak, S. Brânzei, M. Wooldridge, and Nicholas R. Jennings. Implementation and computation of a value for generalized characteristic function games. *ACM Trans. Econ. Comput.*, 2(4):16:1–16:35, 2014.

[Misra *et al.*, 2012] S. Misra, R. Barthwal, and M.S. Obaidat. Community detection in an integrated internet of things and social network architecture. In *GLOBECOM, 2012 IEEE*, pages 1647–1652, 2012.

[Murofushi and Soneda, 1993] T. Murofushi and S. Soneda. Techniques for reading fuzzy measures (iii): Interaction index. In *9th FSS*, pages 693–696, 1993.

[Newman, 2006] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3), 2006.

[Owen, 1972] G. Owen. Multilinear extensions of games. *Management Science*, 18(5-part-2):64–79, 1972.

[Roubens, 1996] M. Roubens. Interaction between criteria and definition of weights in mcdavproblems. In *44th meet. of the Eur. working gr. MCDA*, 1996.

[Shapley, 1953a] L. S. Shapley. A value for n-person games. In *In Contributions to the Theory of Games*, pages 307–317. 1953.

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

[Szczepański *et al.*, 2012] P. L. Szczepański, T.P. Michalak, and T. Rahwan. A new approach to betweenness centrality based on the shapley value. In *AAMAS 2012*, pages 239–246, 2012.

[Szczepański *et al.*, 2014] P.L. Szczepański, T.P. Michalak, and M. Wooldridge. A centrality measure for networks with community structure based on a generalization of the Owen value. In *ECAI'14*, pages 867–872, 2014.

[Szczepański *et al.*, 2015] P.L. Szczepański, M. Tarkowski, T.P. Michalak, P. Harrenstein, and M. Wooldridge. Efficient computation of semivalues for game-theoretic network centrality. In *AAAI'15*, pages 461–469, 2015.

[Zachary, 1977] W.W. Zachary. An information flow model for conflict and fission in small groups. *J. of Anth. Res.*, 33:452–473, 1977.

[Zhou *et al.*, 2009] T. Zhou, L. Lü, and Y.C. Zhang. Predicting missing links via local information. *EPJ B-Condensed Matter and Complex Systems*, 71(4):623–630, 2009.