Closeness Centrality for Networks with Overlapping Community Structure

Mateusz K. Tarkowski[1] and **Piotr Szczepański**[2]

Talal Rahwan[3] and **Tomasz P. Michalak**[1,4] and **Michael Wooldridge**[1]

[2]Department of Computer Science, University of Oxford, United Kingdom
[2] Warsaw University of Technology, Poland
[3]Masdar Institute of Science and Technology, United Arab Emirates
[4]Institute of Informatics, University of Warsaw, Poland

Abstract

Certain real-life networks have a community structure in which communities overlap. For example, a typical bus network includes bus stops (nodes), which belong to one or more bus lines (communities) that often overlap. Clearly, it is important to take this information into account when measuring the *centrality* of a bus stop—how important it is to the functioning of the network. For example, if a certain stop becomes inaccessible, the impact will depend in part on the bus lines that visit it. However, existing centrality measures do not take such information into account. Our aim is to bridge this gap. We begin by developing a new game-theoretic solution concept, which we call the *Configuration semivalue*, in order to have greater flexibility in modelling the community structure compared to previous solution concepts from cooperative game theory. We then use the new concept as a building block to construct the first extension of Closeness centrality to networks with community structure (overlapping or otherwise). Despite the computational complexity inherited from the Configuration semivalue, we show that the corresponding extension of Closeness centrality can be computed in polynomial time. We empirically evaluate this measure and our algorithm that computes it by analysing the Warsaw public transportation network.

Introduction

One of the key problems in network science involves identifying the most important (or central) nodes (Freeman 1979; Dezső and Barabási 2002; Keinan et al. 2004; Page et al. 1999). The four best-known centrality measures are Degree, Betweenness, Closeness and Eigenvector centralities (Bonacich 1972; Freeman 1979), each of which views centrality from a different perspective, focusing on certain traits that make nodes important, or, "central," to the functioning of a network (Brandes and Erlebach 2005; Koschutzki et al. 2005). Our focus in this paper is on Closeness centrality. This measure considers the important nodes to be those that are relatively close to all other nodes in the network: the closer a node is to the others, the higher its centrality. Closeness centrality has many applications, from coauthorship networks (Yan and Ding 2009), through tourism (Shih 2006), to social networks (Barabasi 2003; Karinthy 2006).

One aspect of networks that has been largely ignored in the literature on centrality is the fact that *certain real-life networks have a predefined community structure*. In public transportation networks, for example, bus stops are typically grouped by the bus lines (or routes) that visit them. In coauthorship networks, the various venues where authors publish can be interpreted as communities (Szczepański, Michalak, and Wooldridge 2014). In social networks, individuals grouped by similar interests can be thought of as members of a community. Clearly for such networks, it is desirable to have a centrality measure that accounts for the predefined community structure. Yet, to the best of our knowledge, only one such measure has been developed to date (Szczepański, Michalak, and Wooldridge 2014), which extends Degree centrality to networks with community structure. Despite this recent development, one important aspect of real-life networks remains missing from existing centrality measures: the ability to consider *overlapping communities*. Take social networks, for example, where such overlaps are widespread due to the various affiliations and interests of the individuals involved (Kelley et al. 2012). Likewise, in our example of transportation networks, a bus stop may be on the route of multiple (i.e., overlapping) bus lines. If such a stop becomes inaccessible, then all the bus lines that visit it would no longer function properly. As such, the importance of a bus stop clearly depends (at least partially) on the importance of the bus lines to which it belongs.

In an attempt to define a centrality measure that accounts for overlapping communities, we focused on *game-theoretic centrality measures*.[1] The inspiration behind this line of research comes from solution concepts in *cooperative game theory*. In essence, given a set of *players*, a cooperative solution concept typically defines a payoff for each player by comparing his or her contribution to the various groups of players (more on this in the next section). The rich repository of solution concepts has been extensively refined and expanded over the past decades, making it an ideal toolkit for quantifying the importance of individuals in a setting where those individuals co-exist and operate in groups. In the context of game-theoretic network centrality, the *indi-*

[1]See `www.game-theoretic-centrality.com` and `www.network-centrality.com` for an overview of this line of research.

Solution Concept	Degree	Closeness	Betweenness
Shapley value	[2]	[2]	[3]
Semivalues	[4]	[4]	[3]
Owen value	[5]	This Paper	Open
Coalitional semivalues	[5]	This Paper	Open
Configuration value	Open	This Paper	Open
Configuration semivalues	Open	This Paper	Open

Table 1: *The table outlines the papers that used various solution concepts to extend Degree, Closeness, or Betweenness centralities, and computed them in polynomial time.*

viduals correspond to the *nodes* of the network, while the *groups* correspond to the *subgraphs* of the network. With this mapping, any solution concept from cooperative game theory can be readily applied as a centrality measure, except for one remaining obstacle: computing the solution concept. In fact, most solution concepts are inherently hard to compute (Chalkiadakis, Elkind, and Wooldridge 2011). Fortunately, however, in the context of network centrality, we typically focus on a certain closed-form formula, specifying how each group is evaluated. In certain cases, fixing the group-evaluation function makes it possible to obtain a polynomial-time algorithm that computes the resulting, game-theoretic centrality measure.

With this in mind, we propose the first extension of Closeness centrality to networks with either overlapping or non-overlapping community structure. To this end, we propose four game-theoretic variants of Closeness centrality, three of which are based on existing solution concepts, namely: the *Owen value* (Owen 1977), the *configuration value* (Albizuri, Aurrecoechea, and Zarzuelo 2006), and the *coalitional semivalue* (Szczepański, Michalak, and Wooldridge 2014). The fourth and most general variant is based on a new solution concept proposed in this paper, which generalises the aforementioned three concepts, and offers greater flexibility in modelling the underlying community structure. We call it the *Configuration semivalue*.

Crucially, all of the aforementioned solution concepts are hard to compute given an arbitrary group-evaluation function. Nevertheless, for the purpose of extending Closeness centrality, we propose polynomial-time algorithms for computing the corresponding game-theoretic extension. This result fills several gaps in the computational-complexity analysis of game-theoretic centrality measures (see Table 1).

Finally, to demonstrate the applicability of our approach, we apply it to the Warsaw public transportation network, identifying the most central stops and routes therein, from the perspective of Closeness centrality.

Basic Notation and Definitions

In the following two subsections, we introduce the relevant *game*-theoretic, and *graph*-theoretic concepts, respectively.

[2] Michalak et al. (2013b)

[3] Szczepański, Michalak, and Rahwan (2012, 2016)

[4] Szczepański et al. (2015)

[5] Szczepański, Michalak, and Wooldridge (2014)

Game-Theoretic Concepts

A *cooperative game* consists of a set of players $N = \{1, 2, \dots, n\}$ and a characteristic function $\nu : 2^N \to \mathbb{R}$ such that $\nu(\emptyset) = 0$. This function assigns to each *coalition* of players its payoff (i.e., an indication of its performance). We will henceforth refer to a game simply by ν. A *coalition structure*, $CS = \{Q_1, Q_2, \dots, Q_m\}$, is a partition of N into disjoint coalitions. One of the key questions in cooperative game theory is the following: Given a game ν and a coalition structure CS that the players have formed, how do we divide the payoff of each coalition among its members? In this context, assuming that $CS = \{N\}$, i.e., assuming that the players have formed the *grand coalition*, Shapley (1953) proposed a solution concept—now known as the Shapley value—to fairly divide the payoff from cooperation among the players. Banzhaf (1965) proposed another solution concept—now known as the Banzhaf index—which is similar to the Shapley value except for a subtle difference in the way contributions are weighed. To generalize the aforementioned solution concepts, Weber (1979) proposed *semivalues*—a family of solution concepts that includes both the Shapley value and Banzhaf index. Formally, let $MC(C, i) = \nu(C \cup \{i\}) - \nu(C)$ be the *marginal contribution* of player i to coalition $C \subseteq N \setminus \{i\}$. Denoting by $\beta(k)$ the probability that any player makes a marginal contribution to a coalition of size k, the semivalue of i is:

$$\psi_i(\nu) = \sum_{k=0}^{|N|-1} \beta(k) \mathbb{E}\left[MC(C^k, i)\right], \tag{1}$$

where C^k is a random variable over subsets of size k chosen from the set $N \setminus \{i\}$ with uniform probability, and $\mathbb{E}$ is the expected value operator for this variable. The Shapley value and Banzhaf index are two semivalues, defined by $\beta(k) = 1/|N|$ and $\beta(k) = \binom{|N|-1}{k}/2^{|N|-1}$, respectively.

Importantly, all semivalues assume that $CS = \{N\}$, i.e., that the grand coalition is formed. To relax this assumption, Owen (1977) introduced a solution concept—now known as the Owen value—that divides the payoff of any *a priori* coalition structure CS. Now, when $CS = \{N\}$ or $CS = \{\{i\}\}_{i \in N}$, the Owen value is equivalent to the Shapley value. As such, the Owen value is a generalization of the Shapley value; one that does not generalize the β function (as semivalues do), but rather generalizes the assumed coalition structure CS. Another step in this line of research was taken by Szczepański, Michalak, and Wooldridge (2014), who proposed a generalisation combining both the Owen value and semivalues; they called it *coalitional semivalues*. Formally, given a coalition structure, CS, and discrete probability distributions: $\beta : \{0, \dots, |CS| - 1\} \to [0, 1]$ and $\alpha_j : \{0, \dots, |Q_j| - 1\} \to [0, 1]$ for all $j \in \{1, \dots, |CS|\}$, coalitional semivalues are defined by:

$$\gamma_i(\nu, CS) = \sum_{k=0}^{|CS|-1} \beta(k) \sum_{l=0}^{|Q_j|-1} \alpha_j(l)\, \mathbb{E}\Big[MC\Big(\big(\bigcup T^k\big) \cup C^l, i\Big)\Big] \tag{2}$$

where Q_j is the coalition in CS that player i belongs to, T^k is a random variable over subsets of size k chosen from

$CS \setminus \{Q_j\}$ with uniform probability, C^l is a random variable over subsets of size l chosen from $Q_j \setminus \{i\}$ with uniform probability, and $\mathbb{E}$ is the expected value operator. The coalitional semivalue is equivalent to the Owen value when $\beta(k) = 1/|CS|$ and $\alpha_j(l) = 1/|Q_j|, \forall j \in \{1, \ldots, |CS|\}$.

None of the solution concepts discussed thus far considers overlapping coalitions. To address this issue, Albizuri, Aurrecoechea, and Zarzuelo (2006) generalised the Owen value to situations where the *a priori* coalition structure CS contains overlapping coalitions; they called this generalisation the *Configuration value*. Formally, it is defined as follows, where $\mathcal{T}_i = \{j : j \in \mathbb{N} \text{ and } Q_j \in CS \text{ and } i \in Q_j\}$:

$$\chi_i(\nu, CS) = \sum_{\substack{T \subseteq CS \\ T \cap \mathcal{T}_i = \emptyset}} \sum_{j \in \mathcal{T}_i} \sum_{\substack{C \subseteq Q_j \\ i \notin C}} \lambda MC\Big(\big(\bigcup T\big) \cup C, i\Big),$$

$$\text{where } \lambda = \frac{|T|!(|CS| - |T| - 1)!}{|CS|!} \frac{|C|!(|Q_j| - |C| - 1)!}{|Q_j|!} \quad (3)$$

Graph-Theoretic Concepts

A network is a graph, $G = (V, E)$, comprised of a set of nodes $V = \{v_0, v_1, \ldots, v_{n-1}\}$ and a set of edges $E \subseteq V \times V$. A path is simply a chain of connected nodes. The distance between two nodes, s and t, denoted by $dist(s, t)$ is the length of the shortest path between the two (we assume that $dist(v, v) = 0$). Given a node $v \in V$ and a set of nodes $C \subseteq V$, we say that $dist(C, v)$ is equal to the minimum distance between any node $u \in C$ and v (this implies that if $v \in C$ then $dist(C, v) = 0$).

Closeness centrality quantifies the importance of nodes based on their average distance to other nodes (Freeman 1979). In its most general form, it is formulated as follows:

$$closeness(v) = \sum_{u \in V} f(dist(v, u)),$$

where the function $f : \mathbb{N} \to \mathbb{R}$ determines how the distance influences the centrality. When $f(k) = k$, we obtain the classical Closeness centrality, where the *smaller* the value, the more central the node. In this paper, we focus on an alternative formulation, where $f(k) = 1/k$ and throughout the paper $\frac{1}{0} = 0$. The resulting centrality is known as *harmonic* centrality (Boldi and Vigna 2013). With this modification, the *greater* the value, the more central the node (which is in line with most centrality measures).

Everett and Borgatti (1999) introduced *group Closeness centrality*, which extends the notion of Closeness to groups of nodes as follows:

$$\nu_c(C) = \sum_{u \in (V \setminus C)} f(dist(C, u)). \quad (4)$$

Building upon this formula, the first game-theoretic extension of Closeness centrality was introduced by (Michalak et al. 2013b). In particular, the authors defined a game in which the players are the nodes of the network, and the characteristic function is ν_c. The centrality of each node was then determined using the Shapley value. The resulting game-theoretic centrality measure is called the *Shapley value-based Closeness centrality*. Roughly speaking, the harmonic Closeness centrality evaluates how close a node is to others, whereas the Shapley value-based variant evaluates the role that a node plays in *bringing other nodes closer together*. To illustrate this difference, consider networks **(a)** and **(b)** from Figure 1. Here, according to harmonic Closeness, v_1 is relatively more important in **(b)** than in **(a)** since it is close to more nodes in **(b)** than in **(a)**. In contrast, according to the Shapley-value based Closeness, v_1 is actually more important in **(a)**, since the removal of v_1 from **(a)** has a greater impact on the distances between the other nodes, compared to the removal of v_1 from **(b)**.

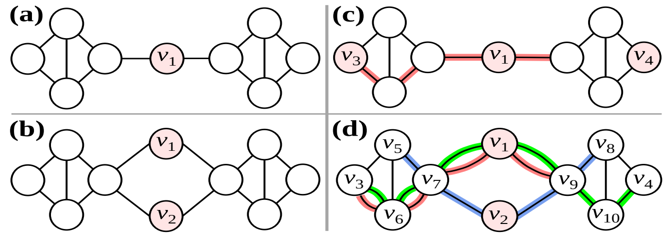

Figure 1: Sample networks. In **(c)** and **(d)**, communities are highlighted by same-coloured edges.

We present in Table 2 the harmonic and Shapley value closeness rankings for Figure 1 (d).[6] The Configuration value closeness ranking is as follows: v_9, v_7, v_1, v_6, v_3, v_8, v_5, v_2, v_{10}, v_4. The configuration value makes use of community information, promoting nodes v_9, v_1, v_6 and v_3. This ranking is also more fine-grained (i.e., there are no ties), because it draws upon community information, which is different for most nodes in the example.

	Harmonic	SV-based
1.	v_7, v_9	v_7, v_9
2.	v_1, v_2	v_1, v_2
3.	v_5, v_6, v_8, v_{10}	v_3, v_4
4.	v_3, v_4	v_5, v_6, v_8, v_{10}

Table 2: *Harmonic closeness and Shapley value rankings for Figure 1 (d).*

Our Centrality Measures

As stated earlier, no centrality measures to date can readily be applied to networks with overlapping community structures. An example is depicted in networks **(c)** and **(d)** from Figure 1. Specifically, in network **(c)**, nodes v_3 and v_4 are symmetric except that v_3 belongs to a seemingly-important community; one that connects the two parts of the network. Arguably, when taking this additional information into consideration, v_3 should be considered more important than v_4. Moving on to network **(d)**, node v_1 belongs to more communities than v_2, and the communities of v_1 seem to be equally important, if not more important, than those of v_2. This could mean, for example, that more bus, tram or train routes visit (and rely on) the bus stop v_1 than the bus stop v_2, implying that v_1 should be more central once the underlying community structure is taken into consideration.

Configuration Semivalues: We now propose a family of solution concepts, which we believe to be the most general

[6]Ranking does not capture the subtlety of valuations. For example, harmonic closeness ranks v_7 as slightly more important than v_1, however Shapley value closeness ranks it as much more important.

of its kind to date, as it not only allows for an arbitrary β, α and CS, but also allows for overlapping coalitions. We call it *Configuration semivalues*, and define it as follows, where T^k is a random variable over subsets of size k of $CS \setminus \mathcal{T}_i$ and $\mathbb{E}$ is the expected value operator:

$$\phi_i(\nu, CS) = \sum_{k=0}^{|CS|-1} \beta(k) \sum_{j \in \mathcal{T}_i} \sum_{l=0}^{|Q_j|-1} \alpha_j(l)\, \mathbb{E}\Big[MC\Big(\big(\bigcup T^k\big) \cup C^l, i\Big)\Big] \tag{5}$$

In particular, given a community structure CS in which communities do not overlap, this family of solution concepts is equivalent to *coalitional semivalues*. Given $CS = \{N\}$, it is equivalent to *semivalues*. Further restrictions on β and α_j (as discussed in the preliminaries) lead to the *Shapley value*, the *Banzhaf index*, or the *Owen value*. Compared to the configuration *value*, our configuration *semivalues* offer greater control over the contributions of players, due to a probability distribution over the number of communities to which a player contributes and a distribution over the number of nodes from his own community that he can contribute to. In applications such as countterrorism (Lindelauf, Hamers, and Husslage 2013; Michalak et al. 2013a), this can represent the expected size of an attack on a network or the number of targeted communities.

Configuration Semivalue Community Index: Whereas to measure the importance of a community using the Owen value it suffices to sum up the power of the nodes comprising it, this is not the case for the Configuration value. In particular, the power of a node may be the result of its membership to many communities. For this reason, the distinction must be made as to which of the player's marginal contributions are made because of which community. To this end, we propose the following measure of community strength:

$$CP_j(\nu, CS) = \sum_{i \in Q_j} \sum_{k=0}^{|CS|-1} \beta(k) \sum_{l=0}^{|Q_j|-1} \alpha_j(l)\, \mathbb{E}\Big[MC\Big(\big(\bigcup T^k\big) \cup C^l, i\Big)\Big],$$

where $C^l \subseteq Q_j \setminus \{i\}$ and $T^k \subseteq CS \setminus Q_j$ are random variables. Although an axiomatic characterisation of this community index is outside the scope of this paper, we mention that in the case of the Configuration value, the index is efficient, and in the case of the Owen value, it is equal to the sum of the Owen values of the members of a community.

Configuration Semivalue Closeness Centrality: Our extension of closeness centrality (which accounts for overlapping and non-overlapping community structures) involves using our Configuration semivalue (see Equation 5) with the characteristic function for group Closeness centrality (see Equation 4). More formally, it is: $\phi_v(\nu_c, CS)$.

Algorithms

We show that any *configuration semivalue* of group Closeness centrality is computable in polynomial time (Theorem 1), and obtain better time complexity for the *configuration value* (Theorem 2).

Theorem 1 *Any configuration* semivalue *of group Closeness centrality for all nodes in a weighted graph $G = (V, E, \omega)$ can be computed in $O(|V|^4|CS|)$ time.*

Proof of Theorem 1: Starting from Equation (5), which defines the configuration semivalue, let us first replace the arbitrary characteristic function therein, i.e., ν, with that of group Closeness centrality (defined in Equation 4). We get:

$$\phi_v(\nu, CS) = \sum_{j \in \mathcal{T}_v} \sum_{k=0}^{|CS|-1} \sum_{l=0}^{|Q_j|-1} \sum_{\substack{T^k \subseteq CS \setminus \mathcal{T}_v \\ |T^k|=k}} \sum_{\substack{C^l \subseteq Q_j \setminus \{v\} \\ |C^l|=l}} \alpha_j(l)\beta(k)$$
$$\frac{\sum_{u \in V} f(dist(\bigcup T^k \cup C^l \cup \{v\}, u)) - f(dist(\bigcup T^k \cup C^l, u))}{\binom{|CS|-1}{k}\binom{|Q_j|-1}{l}}.$$

Although this may seem inconsequential, our next step is to rearrange the summation over u and bring it to the forefront:

$$\phi_v(\nu, CS) = \sum_{u \in V} \sum_{j \in \mathcal{T}_v} \sum_{k=0}^{|CS|-1} \sum_{l=0}^{|Q_j|-1} \sum_{\substack{T^k \subseteq CS \setminus \mathcal{T}_v \\ |T^k|=k}} \sum_{\substack{C^l \subseteq Q_j \setminus \{v\} \\ |C^l|=l}} \alpha_j(l)$$
$$\beta(k) \frac{f(dist(\bigcup T^k \cup C^l \cup \{v\}, u)) - f(dist(\bigcup T^k \cup C^l, u))}{\binom{|CS|-1}{k}\binom{|Q_j|-1}{l}}.$$

Next, for $v \in Q_j$ we will split the equation into:

$$MC^+_{k,l}(v, u, j) = \sum_{\substack{T^k \subseteq CS \setminus \mathcal{T}_v \\ |T^k|=k}} \sum_{\substack{C^l \subseteq Q_j \setminus \{v\} \\ |C^l|=l}} \alpha_j(l)\beta(k)$$
$$\frac{f(dist(\bigcup T^k \cup C^l \cup \{v\}, u))}{\binom{|CS|-1}{k}\binom{|Q_j|-1}{l}}, \text{ and} \tag{6}$$

$$MC^-_{k,l}(v, u, j) = \sum_{\substack{T^k \subseteq CS \setminus \mathcal{T}_v \\ |T^k|=k}} \sum_{\substack{C^l \subseteq Q_j \setminus \{v\} \\ |C^l|=l}} \alpha_j(l)\beta(k)$$
$$\frac{f(dist(\bigcup T^k \cup C^l, u))}{\binom{|CS|-1}{k}\binom{|Q_j|-1}{l}}, \tag{7}$$

with the additional constraint on T^k and C^l such that:

$$dist(\bigcup T^k \cup C^l, u)) \neq dist(\bigcup T^k \cup C^l \cup \{v\}, u)). \tag{8}$$

We can now state the following:

$$\phi_v(\nu, CS) = \sum_{u \in V} \sum_{j \in \mathcal{T}_v} \sum_{k=0}^{|CS|-1} \sum_{l=0}^{|Q_j|-1}$$
$$MC^+_{k,l}(v, u, j) - MC^-_{k,l}(v, u, j). \tag{9}$$

The constraint in Equation (8) simply allows us to avoid redundant computations (the contribution in the opposite case is trivially zero, since by entering such a coalition, v does not change the distance to u). The remainder of the proof will focus on computing Equations (6) and (7). We will first focus on Equation (6).

Note that due to the constraint in Equation (8), we can be sure that $f(dist(\bigcup_{Q \in T^k} Q \cup C^l \cup \{v\}, u)) = f(dist(u, v))$, since by entering the coalition, v must have brought it closer to u. Thus, it suffices to count the number of coalitions from Q_j (of size l)—C^l—and coalitions of communities (of size k)—T^k—such that $dist(\bigcup_{Q \in T^k} Q \cup C^l, u)) > dist(v, u)$, since only then Equation (8) will hold. To do this, let us introduce the following notation:

- $Com_{\sim d}(u) = \{Q : Q \in M \text{ and } dist(Q,u) \sim d\}$;
- $Nod^j_{\sim d}(u) = \{s : s \in C_j \text{ and } dist(s,u) \sim d\}$,

where $\sim$ will be one of $<, >, \leq, \geq$ or $=$. These sets are fairly simple to precompute, and will allow us to count the number of required coalitions. In this particular case, we will use $Com_{>dist(u,v)}$ to count the number of communities farther from u than v. Similarly, $Nod^j_{>dist(u,v)}$ counts the number of nodes in the community Q_j that are farther from u than the distance from u to v. Altogether, there are $\binom{Com_{>dist(u,v)}}{k}$ coalitions of communities and $\binom{Nod^j_{>dist(u,v)}}{l}$ coalitions from Q_j satisfying the requirements. Let $d = dist(u,v)$. Finally:

$$MC^+_{k,l}(v,u,j) = \binom{Nod^j_{>d}}{l}\binom{Com_{>d}}{k} f(dist(v,u)).$$

As for Equation (7), we divide computations as follows:

$$MC^-_{k,l}(v,u,j) = \sum_{d} MC^-_{k,l}(v,u,j,d), \text{ where}$$

$$MC^-_{k,l}(v,u,j,d) = \sum_{\substack{T^k \subseteq CS \setminus \mathcal{T}_v \\ |T^k|=k}} \sum_{\substack{C^l \subseteq Q_j \setminus \{v\} \\ |C^l|=l}} \alpha_j(l)\beta(k) \frac{f(dist(\bigcup T^k \cup C^l, u))}{\binom{|CS|-1}{k}\binom{|Q_j|-1}{l}},$$

keeping in mind the constraint from Equation (8) and adding the constraint on C^l and T^k such that

$$dist(\bigcup T^k \cup C^l, u) = d. \tag{10}$$

Now, the computation of $MC^-_{k,l}(v,u,j,d)$ reduces to computing the number of coalitions C^l and coalitions of communities T^k that satisfy both constraints. By using the inclusion-exclusion principle, and assuming that $dist(v,u) > d$, we have the following:

$$MC^-_{k,l}(v,u,j,d) = \binom{Com_{\geq d}(u)}{k}\binom{Nod^j_{\geq d}(u)}{l} - \binom{Com_{>d}(u)}{k}\binom{Nod^j_{>d}(u)}{l}.$$

To explain this, we first compute the number of coalitions of communities of size k that are at distance d or farther from u. We do the same for coalitions of nodes from Q_j. However, we must take away the number of occurrences when no nodes from C^l and communities from T^k are at distance d from u, in order to satisfy the constraint from Equation (10).

Computing all $MC^-_{k,l}(v,u,j)$ variables is the most time-consuming, as it takes $O(|V|^4|CS|)$ time. This may be counter-intuitive, since computing all $MC^-_{k,l}(v,u,j,d)$ variables would imply $O(|V|^5|CS|)$, but dynamic programming eliminates this need. Precomputations are implemented in Algorithm 1 and marginal contributions in Algorithm 2. $\square$

Theorem 2 *The configuration* value *of group Closeness centrality for all nodes in a weighted graph* $G = (V, E, \omega)$ *can be computed in* $O(|V|^2[\log(|V|) + |CS|] + |V||E|)$ *time.*

Sketch of Proof of Theorem 2: The key idea, is to use an alternate formula for the configuration value:

$$\phi_v(\nu, CS) = \sum_{j \in \mathcal{T}_v} \sum_{\Pi \in \Pi(CS)} \sum_{\pi \in \Pi(Q_j)} \frac{\sum_{u \in V} f(dist(\Pi_{|Q_j} \cup \pi_{|v} \cup \{v\}, u)) - f(dist(\Pi_{|Q_j} \cup \pi_{|v}, u))}{|CS|!|Q_j|!},$$

where $\Pi(X)$ refers to a permutation of the set X, and $\Pi_{|x}$ is the set of elements preceding x in the permutation Π. Next, we group computations for both k and l as follows: $MC^+(v,u,j) = \sum_{k,l} MC^+_{k,l}(v,u,j)$ and $MC^-(v,u,j,d) = \sum_{k,l} MC^-_{k,l}(v,u,j,d)$. By counting the permutations that satisfy the constraints from Theorem 1, we reach the following:

$$MC^+(v,u,j) = \frac{f(d)(|CS|)!(|Q_j|)!}{Nod^j_{\leq d}(u) Com_{\leq d}(u)}, \text{ and}$$

$$MC^-(v,u,j,d) = f(d) \left[\frac{(|CS|)!(|Q_j|)!}{Nod^j_{<d}(u) Com_{<d}(u)} - \frac{(|CS|)!(|Q_j|)!}{Nod^j_{\leq d}(u) Com_{\leq d}(u)}\right]. \square$$

Precomputations are presented in Algorithm 1, and computation of marginal contributions is presented in Algorithm 2. Algorithm 3 computes the Configuration value with better time complexity. In Algorithm 1, lines 1 to 20 compute node distances using Johnson's algorithm (Johnson 1977) and sort them in a descending order, allowing us to use dynamic programming to avoid redundant computation. As an example, if d' directly follows d in such a list, then $COM_{\geq d'}(u) = COM_{=d}(u) + COM_{\geq d}(u)$. The last step removes duplicate distances from the community distances list, which also helps avoid redundant computation.

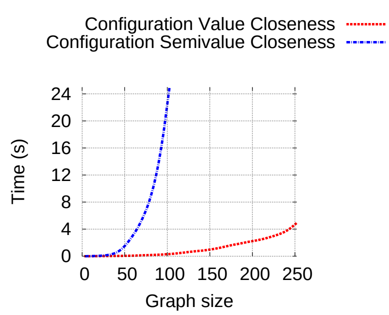

Figure 2: Empirical running times for Algorithms 2 and 3.

Algorithm 2 computes marginal contributions by moving backwards (largest to smallest) through the possible distances, and also uses dynamic programming. The data computed thus-far is held in the variable $prev\_val$ and accumulated in ϕ_v. $s\_merge(a,b)$ is a function that takes two sorted sets (descending order) and returns a sorted set, such that any items in b that are smaller than the smallest item in a are not included. Finally, the array CP must be mentioned, which collects the power of communities as a whole. If a marginal contribution is made by a node through community Q_j, then CP_j is incremented by the value of the contribution.

We conclude by noting that the configuration value is a generalisation of the Shapley value. Algorithm 3 can compute the Shapley value-based Closeness centrality (Michalak et al. 2013b) with the same time-complexity as the best currently known algorithm (computing distances between all nodes limits computation time). Empirical running times are presented in Figure 2.

rank	Harmonic	SV-based	CV-based
1.	Świetokszyska	Politechnika	Centrum
2.	Ratusz	PKP Falenica	Dw.Wileński
3.	Ron.Starzyńskiego	Erazma z Zakr.	PKP Falenica
4.	Dw.Wileński	PKP Radość	Świetokszyska

Table 3: *Top four hubs in the Warsaw transportation network, according to different Closeness measures.*

node	Harmonic	SV	CV	lines
Centrum	234	0.23	4.96	68
Świetokszyska	235	0.50	4.22	29
Politechnika	224	1.04	3.07	17
PKP Falenica	108	1.00	4.91	8
PKP Radość	122	0.80	3.45	8

Table 4: *Properties of hubs in Warsaw.*

Warsaw Public Transportation Network

Algorithms 1, 2 and 3 were implemented in Java. The Configuration-based Closeness centrality (i.e., CV-based Closeness centrality) was used to analyse the Warsaw public transportation network.[7] Edge-weights were defined as the average travel times between nodes. In total, the weighted network consists of 1425 nodes, 2135 edges and 380 communities formed by bus lines, trams, and underground and suburbia trains. We note that the ranking of nodes according to CV-based Closeness differs significantly from the Harmonic Closeness and Shapley value-based (i.e., SV-based) Closeness rankings. We present the four most prominent nodes according to these centralities in Table 3.

Configuration-based Closeness ranks *Centrum* (or "city center") as most important. This result is intuitive and makes sense, since this stop is a large hub in downtown Warsaw, where passengers can choose from 68 bus, tram and train connections, and one metro line (see Table 4). The top ranked nodes according to the other centrality measures are also near the city center, but provide less connections that are not as important as those in the city center.

The most surprising rankings are for *PKP Falenica* and *PKP Radość*, which are railway stations far from downtown Warsaw. They are important because—for certain source-destination pairs—it is almost impossible to find routes that omit these stations. Additionally, trains play an important role in shortening the travel time between distant nodes, since they are the fastest means of transportation.

Interestingly, the fifth node according to CV-based Closeness, *Płowiecka*, is not in the top ten according to the other measures. Harmonic Closeness misses the fact that this stop is not easily replaced. Shapley value-based Closeness misses

[7]The data, experiment results and programs can be downloaded from https://github.com/szczep/gtna.

Algorithm 1: Precomputations for Configuration Semivalue Closeness.

```
input : Graph G = (V, E, ω), Closeness function
        f : ℝ → ℝ, Overlapping Community Structure
        CS, Probability distribution functions
        β : 0, 1, ... |V| − 1 → ℝ,
        ∀_{0≤j≤|CS|−1} α_j : 0, 1, ... |Q_j| − 1 → ℝ
output: Configuration Semivalue
dist[V, V] ← Johnson(G, ω);
for v ∈ V do
    COM[v] ← ∅;
for v ∈ V do
    φ_v ← 0;
    distances[v] ← empty ordered set;
    c_dists[v] ← empty ordered set;
    for Q_j ∈ CS do
        CP[j] ← 0;   com_dist[j][v] ← ∞;
        g_dists[j][v] ← empty ordered set;
        for u ∈ Q_j do
            COM[u] ← COM[u] ∪ Q_j;
            com_dist[j][v] ←
            min(dist(v, u), com_dist[j][v]);
            distances[v] ← distances[v] ∪ ⟨u, dist[u, v]⟩;
        c_dists[v] ← c_dists[v] ∪ ⟨Q_j, com_dist[j][v]⟩;
    sort_desc(c_dists[v]); sort_desc(distances[v]);
for u ∈ V do
    for ⟨v, d⟩ ∈ distances[u] do
        for Q_j ∈ COM[v] do
            g_dists[j][u] ← g_dists[j][u] ∪ ⟨v, d⟩;
for u ∈ V do
    m1 ← largest distance in c_dists[u]; prev ← m1;
    Com_{>prev}(u) ← 0; Com_{≥prev}(u) ← 0;
    for ⟨Q_j, d⟩ ∈ c_dists[u] do
        m2 ← largest distance in g_dists[j][u];
        if m2 > m1 then
            Com_{>m2}(u) ← 0; Com_{≥m2}(u) ← 0;
        if d ≠ prev then
            Com_{>d}(u) ← Com_{≥prev}(u);
            Com_{≥d} ← Com_{≥prev}(u); prev = d;
        Com_{≥d}(u) ← Com_{≥d}(u) + 1;
        prevnod ← largest distance in distances[u];
        Nod^j_{>prevnod}(u) ← 0; Nod^j_{≥prevnod}(u) ← 0;
        for ⟨v, dnod⟩ ∈ distances[u] do
            if dnod ≠ prevnod then
                Nod^j_{>dnod}(u) ← Nod^j_{≥prevnod}(u);
                Nod^j_{≥dnod}(u) ← Nod^j_{≥prevnod}(u);
                prevnod = dnod;
            if Q_j ∈ COM(v) then
                Nod^j_{≥dnod}(u) ← Nod^j_{≥dnod}(u) + 1;
for u ∈ V do
    remove_repeated_distances(c_dists[u]);
```

Algorithm 2: Efficient Algorithm for Configuration Semivalue Closeness (continued from Algorithm 1).

for $u \in V$ **do**
  for $Q_j \in CS, k \in [0, |CS|), l \in [0, |Q_j|)$ **do**
    $prev\_d \leftarrow -1$; $prev\_val \leftarrow 0$; $MC^{-} \leftarrow 0$;
    for $\langle x, d\rangle \in s\_merge(g\_dists[j][u], c\_dists[u])$ **do**
      if $x \in V$ *and* $prev\_d \neq -1$ **then**
        $Com_{\geq d}(u) \leftarrow Com_{\geq prev\_d}(u)$;
        $Com_{>d}(u) \leftarrow Com_{>prev\_d}(u)$;
      if $prev\_d \neq d$ **then**
        $MC^{-} \leftarrow prev\_val$; $prev\_d \leftarrow d$;
        $prev\_val \leftarrow prev\_val + \left[\binom{Com_{\geq d}(u)}{k}\binom{Nod^{j}_{\geq d}(u)}{l} - \binom{Com_{>d}(u)}{k}\binom{Nod^{j}_{>d}(u)}{l}\right] f(d)$;
      if $x \in V$ **then**
        $MC^{+} \leftarrow f(d)\binom{Com_{>d}(u)}{k}\binom{Nod^{j}_{>d}(u)}{l}$;
        $\phi_x \leftarrow \phi_x + \beta(k)\alpha_j(l)\frac{MC^{+}-MC^{-}}{\binom{|CS|-1}{k}\binom{|Q_j|-1}{l}}$;
        $CP_j \leftarrow CP_j + \beta(k)\alpha_j(l)\frac{MC^{+}-MC^{-}}{\binom{|CS|-1}{k}\binom{|Q_j|-1}{l}}$;

Algorithm 3: Efficient Algorithm for Configuration Value Closeness (continued from Algorithm 1).

for $u \in V$ **do**
  for $Q_j \in CS$ **do**
    $prev\_d \leftarrow -1$; $prev\_val \leftarrow 0$; $MC^{-} \leftarrow 0$;
    for $\langle x, d\rangle \in s\_merge(g\_dists[j][u], c\_dists[u])$ **do**
      if $x \in V$ *and* $prev\_d \neq -1$ **then**
        $Com_{\leq d}(u) \leftarrow Com_{\leq prev\_d}(u)$;
        $Com_{<d}(u) \leftarrow Com_{<prev\_d}(u)$;
      if $prev\_d \neq d$ **then**
        $MC^{-} \leftarrow prev\_val$; $prev\_d \leftarrow d$;
        $prev\_val \leftarrow prev\_val + \left[\frac{f(d)}{(Nod^{j}_{<d}(u))(Com_{<d}(u))} - \frac{f(d)}{(Nod^{j}_{\leq d}(u))(Com_{\leq d}(u))}\right]$;
      if $x \in V$ **then**
        $MC^{+} \leftarrow \frac{f(d)}{(Nod^{j}_{\leq d}(u))(Com_{\leq d}(u))}$;
        $\phi_x \leftarrow \phi_x + MC^{+} - MC^{-}$;
        $CP_j \leftarrow CP_j + MC^{+} - MC^{-}$;

that the bus lines that stop at *Płowiecka* are very important. Even if the stop is omitted, a traveller must often still use bus lines that visit *Płowiecka* for further travel. As for communities—unsurprisingly—long routes that bring in commuters from all of Warsaw (including the metro line M1) are ranked as most important.

Harmonic Closeness centrality prioritises the topologically most *central* nodes. The SV-based centrality considers a new dimension, promoting irreplaceable nodes. We go one step further and provide a new CV-based measure that promotes hubs with powerful connections, while taking into account the previous two considerations. Importantly, only our measure is able to detect the most central and popular hub in Warsaw.

Line	Community Index
401	4.35
ZS1	3.68
409	3.61
M1	3.51

Table 5: *Important lines.*

Information Diffusion in Social Networks

Borgatti (2006) advocated the use of group closeness centrality for information diffusion. However, this approach does not yield a comprehensive ranking of nodes, is computationally intractable and does not account for communities. Recently, Lin et al. (2015) noted that communities are important for information diffusion, since intra-community diffusion is much faster than inter-community diffusion.

rank	Harmonic	SV	CV
1.	5274	5274	5
2.	311	35879	823
3.	404	12367	72
4.	727	27977	11287

Table 6: *Top four hubs in the Youtube subnetwork.*

We conducted an experiment on a YouTube social network with ground-truth communities (Mislove et al. 2007) in order to see the impact of overlapping communities on centrality. We chose the 80 first communities from a list of the 5000 most important ones[8] and studied the sub-network consisting of these communities. Figure 6 shows the four most important nodes. Harmonic centrality indicates that node 5274 is topologically most central, whereas SV-based centrality indicates that it also brings other nodes closer together. The configuration value ranking promotes node 5, since it belongs to important communities, and it is important for bringing nodes within these communities closer together. Moreover, it brings other communities closer together, which is imperative for inter-community information transfer.

Conclusions and Future Work

We have developed a general solution concept, namely the Configuration semivalue, that encompasses both coalitional semivalues (Szczepański, Michalak, and Wooldridge 2014) and the configuration value (Albizuri, Aurrecoechea, and Zarzuelo 2006). We have used this value in order to develop the first network centrality measure that accounts for an overlapping community structure. We based our centrality on the notion of Closeness, developed polynomial-time algorithms for its computation and used it to analyse the Warsaw public transportation network. This research also fills a gap in the complexity analysis of game-theoretic centrality measures. An interesting direction for future work is to complete the missing entries in Table 1. Finally, although the configuration value has been axiomatised, configuration semivalues in general and community indices need further study.

[8]Available at https://snap.stanford.edu/data/com-Youtube.html

Acknowledgements

Tomasz Michalak and Michael Wooldridge were supported by the European Research Council under Advanced Grant 291528 ("RACE"). This work was also supported by the Polish National Science Centre grant DEC-2013/09/D/ST6/03920.

References

Albizuri, M.; Aurrecoechea, J.; and Zarzuelo, J. 2006. Configuration values: Extensions of the coalitional owen value. *GAME ECON BEHAV* 57(1):1 – 17.

Banzhaf, J. F. 1965. Weighted Voting Doesn't Work: A Mathematical Analysis. *RUTGERS LAW REV* 19:317–343.

Barabasi, A.-L. 2003. *Linked: How Everything Is Connected to Everything Else and What It Means*. Plume.

Boldi, P., and Vigna, S. 2013. Axioms for centrality. *CoRR* abs/1308.2140.

Bonacich, P. 1972. Factoring and weighting approaches to status scores and clique identification. *J MATH SOCIOL* 2(1):113–120.

Borgatti, S. 2006. Identifying sets of key players in a social network. *COMPUT MQTH ORGAN TH* 12(1):21–34.

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

Chalkiadakis, G.; Elkind, E.; and Wooldridge, M. 2011. *Computational aspects of cooperative game theory*. Morgan & Claypool Publishers.

Dezső, Z., and Barabási, A.-L. 2002. Halting viruses in scale-free networks. *PHYS REV E* 65:055103.

Everett, M. G., and Borgatti, S. P. 1999. The centrality of groups and classes. *J MATH SOCIOL* 23(3):181–201.

Freeman, L. 1979. Centrality in social networks: Conceptual clarification. *SOC NETWORKS* 1(3):215–239.

Johnson, D. B. 1977. Efficient algorithms for shortest paths in sparse networks. *J ACM* 24(1):1–13.

Karinthy, F. 2006. Chain-links. In M. Newman, A.-L. B., and Watts, D., eds., *The Structure and Dynamics of Networks*. Princeton University Press. 21–26.

Keinan, A.; Hilgetag, C. C.; Meilijson, I.; and Ruppin, E. 2004. Casual localization of neural function: the shapley value method. *NEUROCOMPUTING* 58-60(0):215–222.

Kelley, S.; Goldberg, M.; Magdon-Ismail, M.; Mertsalov, K.; and Wallace, A. 2012. Defining and discovering communities in social networks. In Thai, M. T., and Pardalos, P. M., eds., *Handbook of Optimization in Complex Networks*, volume 57 of *Springer Optimization and Its Applications*. Springer US. 139–168.

Koschutzki, D.; Lehmann, K.; Peeters, L.; Richter, S.; Tenfelde-Podehl, D.; and Zlotowski, O. 2005. Centrality indices. In Brandes, U., and Erlebach, T., eds., *Network Analysis*, volume 3418 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 16–61.

Lin, S.; Hu, Q.; Wang, G.; and Yu, P. 2015. Understanding community effects on information diffusion. In Cao, T.; Lim, E.-P.; Zhou, Z.-H.; Ho, T.-B.; Cheung, D.; and Motoda, H., eds., *Advances in Knowledge Discovery and Data Mining*, volume 9077 of *Lecture Notes in Computer Science*. Springer International Publishing. 82–95.

Lindelauf, R.; Hamers, H.; and Husslage, B. 2013. Cooperative game theoretic centrality analysis of terrorist networks: The cases of jemaah islamiyah and al qaeda. *EUR J OPER RES* 229(1):230–238.

Michalak, T. P.; Rahwan, T.; Szczepański, P. L.; Skibski, O.; Narayanam, R.; Wooldridge, M. J.; and Jennings, N. R. 2013a. Computational analysis of connectivity games with applications to the investigation of terrorist networks. *IJCAI'13*.

Michalak, T.; Aaditha, K. V.; Szczepański, P. L.; Ravindran, B.; and Jennings, N. R. 2013b. Efficient computation of the shapley value for game-theoretic network centrality. *JAIR* 46:607–650.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2007. Measurement and Analysis of Online Social Networks. In *IMC'07*.

Owen, G. 1977. Values of games with a priori unions. In Henn, R., and Moeschlin, O., eds., *Mathematical Economics and Game Theory*, volume 141 of *Lecture Notes in Economics and Mathematical Systems*. Springer Berlin Heidelberg. 76–88.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab.

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

Shih, H.-Y. 2006. Network characteristics of drive tourism destinations: An application of network analysis in tourism. *TOURISM MANAGE* 27(5):1029 – 1039.

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

Szczepański, P. L.; Michalak, T.; and Rahwan, T. 2012. A new approach to betweenness centrality based on the shapley value. In *AAMAS'12*, 239–246.

Szczepański, P. L.; Michalak, T. P.; and Rahwan, T. 2016. Efficient algorithms for game-theoretic betweenness centrality. *ARTIFICIAL INTELLIGENCE* 231:39 – 63.

Szczepański, P. L.; Michalak, T. P.; and Wooldridge, M. 2014. A centrality measure for networks with community structure based on a generalization of the owen value. In *ECAI*.

Weber, R. J. 1979. Subjectivity in the Valuation of Games. Cowles Foundation Discussion Papers 515, Cowles Foundation for Research in Economics, Yale University.

Yan, E., and Ding, Y. 2009. Applying centrality measures to impact analysis: A coauthorship network analysis. *J AM SOC INF SCI TEC* 60(10):2107–2118.