Mitri Kitti, Matti Pihlava, and Hannu Salonen

Recursive Clustering Methods for Network Analysis

Aboa Centre for Economics

Discussion paper No. 118

Turku 2018

The Aboa Centre for Economics is a joint initiative of the economics departments of the University of Turku and Åbo Akademi University.

ISSN 1796-3133

Printed in Uniprint
Turku 2018

Mitri Kitti, Matti Pihlava, and Hannu Salonen

Recursive Clustering Methods for Network Analysis

Aboa Centre for Economics
Discussion paper No. 118
January 2018

ABSTRACT

We study axiomatically recursive clustering methods for networks. Such methods can be used to identify community structures of a network. One of the methods is based on identifying a node subset that maximizes the average degree within this subset. Once such a subset is found, the method is applied on the subnetwork whose node set is the complement of the first cluster, and so on recursively. The method produces an ordered partition of the node set of the original network. We give a list of axioms that this method satisfies, and show that any recursive clustering method satisfying the same set of axioms must produce the same or a coarser partition than our method.

JEL Classification: C71; D85

Keywords: networks; clustering; community structure

Contact information

Hannu Salonen
Department of Economics
University of Turku
FI-20014, Finland
Email: hannu.salonen (at) utu.fi

Acknowledgements

We thank the Academy of Finland and the OP-Pohjola Group Research Foundation for financial support.

1. Introduction

We study axiomatically recursive clustering methods, or "clustering functions", for weighted directed networks. These methods can be used to identify community structures of such networks. Axiomatic analysis of this subject is important because it forces researchers to think carefully what in their opinion constitutes a community within a network.

A clustering function F defined and studied in this paper is based on identifying a node subset that maximizes the average degree within this subset. Once such a subset is found, the method is applied on the subnetwork whose node set is the complement of the first cluster (the "main cluster"), and so on recursively. This method produces typically an ordered partition of the node set of the original network. The partition is interpreted as the community structure this network. The recursive clustering methods studied here can be viewed as belonging to the class of "divisive hierarchical methods" (see Fortunato 2010).

The function F satisfies axioms such as connectedness (clusters are connected subsets), independence of irrelevant alternatives (reducing the weights of links outside the main cluster does not affect the main cluster), and the average degree monotonicity (the average degree within the main cluster cannot be less than then the average degree of the whole network). We show that if f is any clustering function satisfying these axioms, then f generates the same partition or a coarser partition than F.

Kleinberg (2002) studies clustering methods from the axiomatic point of view. Given a finite set X and any distance function d on X, Kleinberg's clustering function assigns to each d a clustering of X. Distance function d satisfies the usual properties of a metric except that the triangle inequality need not be satisfied. He shows that there exists no clustering method that satisfies three axioms: scale invariance, richness, and consistency. Scale invariance says that clustering does not change if the distance function is multiplied by a positive constant. Richness says that given any partition of a finite set, there is some distance function that gives the members of that partition as clusters. Kleinberg's consistency axiom resembles the independence of irrelevant alternatives axiom.

There are ways to overcome this impossibility result. Ackerman and Ben-David (2008) observed that the axioms are consistent when applied on the set of clustering quality measures (sometimes called clustering quality functions). Such a measure assigns a non-negative real number to every clustering C of the space (X, d). Ackerman and Ben-David (2008) show that there are several clustering quality measures that satisfy the axioms of Kleinberg.

There may be other ways to resolve the impossibility result, since Kleinberg's model does not cover all network models. In weighted networks, if a link between two nodes has a large weight, the interpretation is that the connection between these nodes is very strong, or that these nodes are very similar.

On the other hand, a large value of distance $d(i, j)$ means that nodes i, j are far apart of each other, or that these nodes are very dissimilar. That difference in interpretation need not be of importance: since the distance has properties $d(i, j) > 0$ if $i \neq j$, and $d(i, j) = d(j, i)$ for all $i, j \in X$. But Kleinberg's model does not cover all weighted or directed networks since link weight may be zero between two distinct nodes and symmetry is not necessarily satisfied (see Carlsson *et.al.* (2014) for an axiomatic analysis of hierarchical clustering methods when $d(i, j) = d(j, i)$ need not hold).

van Laarhoven and Marchiori (2014) study axiomatically clustering functions and clustering quality functions on the set of weighted undirected networks, and show that Kleinberg's axiom are consistent. In fact, a clustering function that assigns to each network the partition into connected components satisfies all the axioms. van Laarhoven and Marchiori (2014) give also a list of reasonable axioms that any clustering quality function should satisfy and note that the modularity does not satisfy all these axioms but a version of it does (for modularity, see Newman and Girwan 2004).

The literature of clustering methods and community detection is growing rapidly. For useful reviews, see for example Fortunato (2010) and Luxburg (2007).

The paper is organized in the following way. In Section 2 notation, definitions and axioms are given. In Section 3 the results are stated and proved.

2. Preliminaries

Let $G = (V, A)$ denote a *weighted network*. Here V is a nonempty set of nodes that are typically indexed by natural numbers, say $V = \{1, \ldots, n\}$, and A is a nonnegative matrix representing the link structure. Denote by a_{ij} the element of A that lies on i'th row and j'th column. If the network is *undirected*, then the matrix A is symmetric: $a_{ij} = a_{ji}$ for all $i, j \in V$. If A is not symmetric, then G is *directed*, and the element a_{ij} is interpreted as the strength of the directed link from i to j. If $a_{ij} = 0$ we interpret that there is no link between i and j. In this paper we assume $a_{ii} = 0$ for all i, so we ignore loops.

Let $\mathcal{G}$ denote the set of all weighted networks, and let $\mathcal{G}^u$ denote the set of all undirected weighted networks. Given $G = (V, A) \in \mathcal{G}$, $G = (V, A)$ is a subnetwork of $G = (V, A)$, if $V' \subset V$ and A' is the restriction of the matrix A to the node set V'.

An undirected network G can be partitioned into connected components: $V = V_1 \cup \cdots \cup V_m$ such that each V_k is nonempty and connected, and $v \notin V_k$ implies $a_{iv} = 0 = a_{vi}$ for all $i \in V_k$. Connectedness of V_k means that for each $i, j \in V_k$ there exists a path in V_k: there are nodes $i = i_0, \ldots, i_t = j$ in V_k such that there is link between i_m and i_{m+1} for each $m < t$. If V is the only component, then G is connected. This happens if A is *irreducible*: for each $i, j \in V$ there is some $t \in \mathbb{N}$ such that $a^t_{ij} > 0$, where a^t_{ij} is the ij -element of the matrix A^t. A directed network can also be partitioned into connected components if the directions of links are ignored.

If for each $i, j \in V_k$ there is a directed path from i to j and from j to i such that the whole path lies in V_k, then V_k is *strongly connected*. The node set V of $G = (V, A)$ can be partitioned into strongly connected components. Singletons $\{v\}$ are interpreted as members of that partition, if it holds that $\{v\} \cup X$ is not strongly connected for any strongly connected subset $X \subset V$ such that $|X| \geq 2$.

Denote by $P(G)$ the set of all nonempty partitions of the node set V of G. If there is no possibility for mistakes about what G is, we may denote by $P(V)$ the set of all partitions of V.

A function f defined by $G \longrightarrow f(G) \in P(G)$ is called a clustering function. In this paper we study recursive clustering functions f. These functions generate an ordered partition $f(G) = \langle C_1, \ldots, C_m \rangle$ in the following way. First f selects a nonempty subset $C_1 \subset V$ as the *main cluster*. If $C_1 \neq V$, then f selects a subset $C_2 \subset V \setminus C_1$ as the main cluster of the subnetwork $G' = (N \setminus C_1, A')$, where A' is the submatrix of A corresponding to the node set $V \setminus C_1$. The recursion is continued until a subnetwork is found such that the whole node set $V \setminus (C_1 \cup \cdots \cup C_{m-1})$ is its main cluster C_m. It may happen that some $C_k \in f(G)$ is a collection of pairwise disjoint subsets instead of being a unique subset. In that case it should be understood that the ordering of $f(G)$ is over the collections $C_k \in f(G)$.

We adopt the notation $\sum_{S,T} a_{ij} \equiv \sum_{i \in S} \sum_{j \in T} a_{ij}$, for any $S, T \subset V$. We may denote the complement of S either by S^c or by $V \setminus S$. The cardinality of S is denoted by $|S|$.

2.1. Axioms

Our first axiom says that the members C of the partition $f(G)$ are connected subsets. Strong connectedness seems to be overly restrictive property for clusters. Take for example a directed star network on $V = \{1, 2, 3\}$ such that the node 1 is the center of the star, and the only links are $2 \to 1$ and $3 \to 1$. This network is connected since V is a connected component, and the only natural cluster is $V = \{1, 2, 3\}$. But this network is not strongly connected since the partition into strongly connected components is $\{\{1\}, \{2\}, \{3\}\}$.

Axiom 1 (Connectedness)**.** The members C of the partition $f(G)$ are connected subsets, for all networks G.

Connectedness implies that if $C \in f(G)$ then C must be a subset of some connected component of G. The following axiom is familiar from many different contexts. Here it says that if a network is modified so that the link strengths inside the main cluster do not change, but other links may get weaker, then the original main cluster is the main cluster of the modified network as well.

Axiom 2 (Independence of Irrelevant Alternatives)**.** Given $G = (V, A)$ and its main cluster $C_1 \in f(G)$, let $G' = (V', A')$ be such that (1) $C_1 \subset V' \subset V$;

(2) A' is the matrix A restricted to V' such that $a'_{ij} = a_{ij}$ for every $i, j \in C_1$, and $a'_{ij} \leq a_{ij}$ for every i, j such that $i \notin C_1$ or $j \notin C_1$. Then $C_1 \in f(G')$ is the main cluster of G' as well.

The following axiom states a necessary condition for a proper subset $S \subset V$ to be a main cluster of $G = (V, A)$. It says that the average degree within S must not be less than the average degree in the whole network.

Axiom 3 (Average Degree Monotonicity)**.** Given a network $G = (V, A) \in \mathcal{G}$, if $C_1 \in f(G)$ is the main cluster of G, and $C_1 \neq V$, then

$$\frac{\sum_{S,S} a_{ij}}{|S|} \geq \frac{\sum_{V,V} a_{ij}}{|V|},$$

the inequality being strict if V is connected.

This axiom is motivated by the following example.

Example 1. Take a directed network $G = (V, A)$ such that $V = \{1, 2, 3, 4\}$, $a_{12} = a_{21} = 1$, $a_{34} = a_{43} = 1$, and $a_{ij} = 0$ for all other nodes. This network is not connected and any cluster must be a subset of $\{1, 2\}$ or $\{3, 4\}$. The only reasonable partition is $\{\{1, 2\}, \{3, 4\}\}$, and the average degrees of these subsets are the same as the average degree of the whole network. $\lhd$

2.2. Clustering function F.

Given any nonempty $S \subset V$, let define $v(S)$ by:

$$v(S) = \frac{\sum_{S,S} a_{ij}}{|S|}.$$

Let

$$\begin{aligned} &C_1 \in \mathrm{argmax}_S\{v(S) \mid S \subset V,\ S \ \text{ is connected}\},\ \textit{and} \qquad (1)\\ &C_k \in \mathrm{argmax}\{v(S) \mid S \subset V \setminus (C_1 \cup \cdots \cup C_{k-1}), S \text{ is connected}\}, \end{aligned}$$

with the qualification that if at any stage $k = 1, \ldots$ there are two (or more) subsets S_1, S_2 that maximize v, and $S_1 \cup S_2$ is a connected subset that also maximizes v, then take their union. If on the other hand there are two (or more) subsets S_1, S_2 that maximize v, but $S_1 \cup S_2$ is not connected, then take both S_1 and S_2 as members of the partition at that stage k. In this case it is understood that C_k is a collection of disjoint subsets. In the next section we will show that if $S_1 \cup S_2$ is not connected, then both S_1 and S_2 cannot maximize v.

In words, look at node subsets S and compute the sum of link strengths within this subset, divided by the number of nodes in S, to get value $v(S)$ (sometimes called the density of S). Take as the first element of the partition the subset C_1 having the greatest $v(S)$, etc recursively, with the qualifications given about how to handle the cases with many maximizers. In other words C_1 is the subset such that the average degree (within C_1) reaches its maximum at C_1 among all subsets S of V.

Denote by F the function that assigns such a partition to any network. We will prove in the next section that F is a function, *i.e.*, it generates a unique partition to each G.

3. Results

Our first result states that F produces a unique partition.

Proposition 1. *The mapping F is single valued: $F(G)$ is a unique partition for each $G \in \mathcal{G}$.*

Proof. See the Appendix. □

The following proposition says that the function F satisfies all the axioms presented in the previous Section.

Proposition 2. *The recursive clustering function F satisfies axioms 1, 2, and 3.*

Proof. See the Appendix. □

The following lemma will be useful in the proof of the main result.

Lemma 1. *If a recursive clustering function f satisfies axioms 1, 2, and 3 on the set of directed weighted networks $\mathcal{G}$, then for any $G = (V, A) \in \mathcal{G}$, to any main cluster $C(f)$ generated by f for G there exists a cluster C generated by F for G such that $C \subset C(f)$.*

Proof. Suppose there exists a network $G = (V, A)$ and a main cluster $C(f)$ generated by f for G such that $C \subset C(f)$ does not hold for any cluster C generated by F for G.

If there exists a main cluster $C(F)$ generated by F for G such that $C(F) \cup C(f)$ is connected, then consider a network $G' = (V', A')$ such that $V' = C(F) \cup C(f)$ and matrix A' is the restriction of A to V'. By the axiom 2, independence of irrelevant alternatives, the main clusters generated by F and f for G' are $C(F)$ and $C(f)$, respectively.

By the definition of F and equation (1), $v(C(F)) > v(C(f))$. But then $v(V') = v(C(F) \cup C(f)) \geq v(C(f))$ (see the proof of Proposition 1), a contradiction with axiom 3, average degree monotonicity. Hence $C(F) \subset C(f)$ holds, a contradiction.

Therefore, if $C \subset C(f)$ does not hold for any cluster C generated by F for G, then $C(F) \cup C(f)$ cannot be connected. Then take the least index i such that $C_i \cup C(f)$ is connected for $C_i \in F(G) = \langle C_1, \ldots, C_k \rangle$. (If C_i is a collection instead of a single subset, interpret C_i as being any member of this collection.)

Suppose that $C_i \neq C(f)$, and let $G' = (V', A')$ be a network such that $V' = C_i \cup C(f)$ and A' is the matrix A restricted to V'. By the axiom 2, independence of irrelevant alternatives, the main cluster generated by f for G' is $C(f)$.

By the definition of F and equation (1), $v(C_i) > v(V')$, and hence C_i is a main cluster generated by F for G'. Then by the same argument that was used above for the case when $C(F) \cup C(f)$ is connected, we get that $C_i \subset C(f)$. □

The following theorem says that any recursive clustering function satisfying axioms 1, 2, and 3 will produce the same partition as F or a coarser partition than F. It should be noted that this theorem does not say that the *ordered* partition $f(G)$ is the same as the ordered partition $F(G)$.

Theorem 1. *If a recursive clustering function f satisfies axioms 1, 2, and 3 on the set of all directed weighted networks $\mathcal{G}$, then f generates the same partition as F or a coarser partition than F, for any $G \in \mathcal{G}$.*

Proof. Let then f be any recursive clustering function satisfying axioms 1, 2, and 3, and let $G = (V, A)$ be any directed weighted network.

By Lemma 1, to any main cluster $C(f)$ generated by f for G, there exists a cluster $C \in F(G)$ such that $C \subset C(f)$. Let $C' \in F(G)$ be any cluster such that $C' \cap C(f) \neq \emptyset$. By the same argument that was used in the proof of Lemma 1, we can conclude that $C' \subset C(f)$. Hence $C(f)$ is a union of some clusters C generated by F for G.

Consider a network $G' = (V', A')$ such that $V' = V \setminus C(f)$ for a main cluster $C(f)$ generated by f for G. The matrix A' is the restriction of A to V'.

If V' contains a main cluster C^* generated by f for G, then by independence of irrelevant alternatives C^* must be a main cluster generated by f for G' as well. By the definition of F, every cluster $C \in F(G)$ such that $C \subset V'$ must be a cluster generated by F for G' as well. Hence C^* is a union of some clusters C generated by F for G'.

If V' does not contai any main cluster C^* generated by f for G, then since f is a recursive clustering function, the main cluster(s) $C'(f)$ generated by f for G' must be clusters of the original network G. It follows by Lemma 1 that for any main cluster $C'(f)$ generated by f for G', there exists a cluster $C' \in F(G')$ such that $C' \subset C'(f)$. The proof is completed by induction. □

Remark 1. Given a network $G = (V, A)$, index the nodes in V by natural numbers $1, \ldots, n$ where $n = |V|$. Modify the function F so that it chooses as the main cluster C_1 that subset S that maximizes v in equation (1) and that contains the least index i if there are several maximizers. This function will generate an ordered partition $\langle C_1^*, \ldots, C_m^* \rangle$ such that each C_i^* is a unique subset, and each C_i^* is a cluster in $F(G)$ as well. ◁

Given a network G, let $\mathcal{P}^F(G) = \{f(G) \mid f \text{ satisfies the axioms } 1, 2, 3\}$. So $\mathcal{P}^F(G)$ is the set of all partitions generated for G by at least one of the functions satisfying axioms $1, 2,$ and 3. The following proposition says that the axiom system in Theorem 1 is independent in the sense that if any one of the three axioms is not assumed, then then there is clustering function f and a network G such that $f(G) \notin \mathcal{P}^F(G)$

Proposition 3. *If any one of the axioms 1, 2, and 3 is not assumed, then there is a clustering function f satisfying the other two axioms such that that $f(G) \notin \mathcal{P}^F(G)$, for some $G = (V, A)$.*

Proof. Given $G = (V, A)$, and $F(G) = \langle C_1, \ldots, C_k \rangle$, let $f(G)) = \langle C_1^*, \ldots, C_k^* \rangle$ such that C_i^* is the union of subsets in C_i, $i = 1, \ldots, k$. This f satisfies all the other axioms except axiom 1, connectedness, and $f(G) \notin \mathcal{P}^F(G)$ for some networks G.

Let $f(G)$ be the partition of V into connected components. This f satisfies all the other axioms except axiom 3, average degree monotonicity, and $f(G) \notin \mathcal{P}^F(G)$ for some networks G.

Given $G = (V, A)$, take a connected component $K \subset V$ if V is not connected. Index nodes of K by natural numbers $1, \ldots, k$, $k = |K|$. Choose as the main cluster a connected subset $C_1 \subset K$ such that $1 \in C_1$ and C_1 maximizes v given by equation (1). Let $G' = (V', A')$, where $V' = K \setminus C_1$ and A' is the matrix A restricted to V'. Take a connected component K' of V' if V' is not connected. Index K' by natural numbers $1, \ldots, k'$, where $k' = |K'|$. Choose as the main cluster a subset $C_2 \subset K'$ such that $1 \in C_2$ and C_2 maximizes v given by equation (1).

Continue in this fashion until the node set K is exhausted, and construct an ordered partition $\langle C_1, \ldots, C_m \rangle$. This method satisfies the other axioms except independence of irrelevant alternatives, and $f(G) \notin \mathcal{P}^F(G)$ for some networks G. □

Appendix

Proof of Proposition 1. Assume that at some stage of the equation (1) defining F, there are several maximizers of $v(S)$. Let S and T be two of them. Then

$$\frac{\sum_{S,S} a_{ij}}{|S|} = \frac{\sum_{T,T} a_{ij}}{|T|}.$$

Assume first that $S \cap T = \emptyset$. Consider the union $S \cup T$, and note that $v(S \cup T) \leq v(S)$ must hold, that is

$$\frac{\sum_{S,S} a_{ij} + \sum_{T,T} a_{ij} + \sum_{S,T} a_{ij} + \sum_{T,S} a_{ij}}{|S| + |T|} \leq \frac{\sum_{S,S} a_{ij}}{|S|}, \tag{2}$$

since S maximizes $v(S)$. After a couple of straightforward simplifying steps we get that

$$\frac{\sum_{T,T} a_{ij}}{|T|} + \frac{\sum_{S,T} a_{ij} + \sum_{T,S} a_{ij}}{|T|} \leq \frac{\sum_{S,S} a_{ij}}{|S|}. \tag{3}$$

This inequality holds only if $\sum_{S,T} a_{ij} = 0 = \sum_{T,S} a_{ij}$, and then it holds as an equality. In this case $S \cup T$ is not connected, and then by the definition of F we take both S and T as members of the partition.

If $\sum_{S,T} a_{ij} > 0$ or $\sum_{T,S} a_{ij} > 0$ then inequality (3) cannot hold, a contradiction with the assumption that both S and T maximize $v(S)$. That means that if $v(S) = v(T)$, $S \cap T = \emptyset$ and $S \cup T$ is connected, then $v(S \cup T) > v(S)$.

Suppose next that $S \cap T \neq \emptyset$. Note that $S \cup T$ is necessarily connected. Then $v(S \cup T) \leq v(S)$ means that

$$\frac{\sum_{S,S} a_{ij} + \sum_{T,T} a_{ij} + \sum_{S\setminus T,T\setminus S} a_{ij} + \sum_{T\setminus S,S\setminus T} a_{ij} - \sum_{S\cap T,S\cap T} a_{ij}}{|S| + |T| - |S \cap T|} \leq \frac{\sum_{S,S} a_{ij}}{|S|}.$$

After a couple of straightforward simplifying steps we get that

$$\frac{\sum_{S,S} a_{ij}}{|S|} + \frac{\sum_{S\setminus T, T\setminus S} a_{ij} + \sum_{T\setminus S, S\setminus T} a_{ij}}{|S \cap T|} \leq \frac{\sum_{S\cap T,S\cap T} a_{ij}}{|S \cap T|}. \tag{4}$$

But since T maximizes $v(T)$, this inequality can hold only if

$$\sum_{S\setminus T, T\setminus S} a_{ij} = 0 = \sum_{T\setminus S, S\setminus T} a_{ij},$$

and then equation (4) holds as an equality. That means that also $S \cap T$ maximizes $v(S)$, and in fact it can be shown by a simple example that this is possible. By the definition of F, $S \cup T$ is chosen as a member of the partition.

If there are several subsets $S_1, \ldots, S_k$ maximizing $v(S)$ such that their union is connected, then the previous proof can be applied and the union $S_1 \cup \cdots \cup S_k$ should be chosen as a member of the partition. □

Proof of Proposition 2. Assume first that the maximizers C_k in equation (1) are unique subsets instead of a collections of subsets. Then the axioms 1, 2, and 3 (connectedness, independence of irrelevant alternatives, and average degree monotonicity) are satisfied by definition of F.

Assume then that C_1 and/or some other C_k may consist of several clusters. Then if $A, B \in C_k$ and $i \in A, j \in B$, it holds that $v(A) = v(B)$, and

$a_{ij} = a_{ji} = 0$ (see the proof of Proposition 1). It is clear that F satisfies the axioms also in this case. □

References

Ackerman, M. and Ben-David, S. (2008) Measures of Clustering Quality: A Working Set of Axioms for Clustering. In *Advances in Neural Information Processing Systems* (NIPS), Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds.), pp. 121–128, Curran Associates, Inc.

Carlsson, G., Memoli, F., Ribeiro, A., and Segarra, S. (2014) Axiomatic Construction of Hierarchical Clustering in Asymmetric Networks. *arXiv:1301.7724v2 [cs.LG] 2*, Sep 2014.

Fortunato, S. (2010) Community detection in graphs. *Physich Reports* **486**: 75–174.

Kleinberg, J. (2002) An impossibility theorem for clustering. In *Advances in Neural Information Processing Systems* (NIPS), Becker, S., Thrun, S., and Obermayer, K. (eds.), pp. 446–453, MIT Press.

von Luxburg, U. (2007) A tutorial on spectral clustering. *Statistics and Computing* **17**: 395–416.

Newman.M.E.J., and Girvan, M. (2004) Finding and evaluating community structure in networks. *Physical Review E*: 69:026113, doi: 10.1103/PhysRevE.69.026113.

van Laarhoven, T., and Marchiori, E. (2014) Axioms for Graph Clustering Quality Functions. *Journal of Machine Learning Research* **15**: 193–215.

The **Aboa Centre for Economics (ACE)** is a joint initiative of the economics departments of the Turku School of Economics at the University of Turku and the School of Business and Economics at Åbo Akademi University. ACE was founded in 1998. The aim of the Centre is to coordinate research and education related to economics.

Contact information: Aboa Centre for Economics, Department of Economics, Rehtorinpellonkatu 3, FI-20500 Turku, Finland.

www.ace-economics.fi

ISSN 1796-3133