Social Networks 42 (2015) 80–87

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Generalized two-mode cores

Monika Cerinšek[a,*], Vladimir Batagelj[b,c]

[a] Hruška d.o.o., Kajuhova 90, 1000 Ljubljana, Slovenia
[b] Institute of Mathematics, Physics and Mechanics, Jadranska 19, 1000 Ljubljana, Slovenia
[c] University of Ljubljana, FMF, Department of Mathematics, Jadranska 19, 1000 Ljubljana, Slovenia

ARTICLE INFO

Keywords:
Network analysis
Two-mode network
Generalized two-mode core
Algorithm

ABSTRACT

The node set of a two-mode network consists of two disjoint subsets and all its links are linking these two subsets. The links can be weighted. We developed a new method for identifying important subnetworks in two-mode networks. The method combines and extends the ideas from generalized cores in one-mode networks and from (p, q)-cores for two-mode networks. In this paper we introduce the notion of generalized two-mode cores and discuss some of their properties. An efficient algorithm to determine generalized two-mode cores and an analysis of its complexity are also presented. For illustration some results obtained in analyses of real-life data are presented.

1. Introduction

Network analysis is an approach to the analysis of relational data. In this paper we deal with the analysis of two-mode networks (Borgatti and Halgin, 2011; Latapy et al., 2008). A two-mode network is a network in which the set of nodes consists of two disjoint subsets and its links are linking these two subsets.

The traditional approach to the analysis of two-mode networks is usually indirect: first a two-mode network is converted into one of the two corresponding one-mode projections, and afterward it is analyzed using standard network analysis methods (Wasserman, 1994). *Direct* methods for the analysis of two-mode networks are quite rare (Agneessens and Everett, 2013; Batagelj, 2009; Scott and Carrington, 2011). We can use bipartite statistics on degrees, generalized blockmodeling, (p, q)-cores, two-mode hubs and authorities, 4-ring weights, bi-communities, two-mode clustering, bipartite cores, and some others. Many methods for identifying important subnetworks are available for one-mode networks (measures of centrality and importance, generalized cores, line islands, node islands, clustering, blockmodeling, etc.). We present a new direct method which can be used for the identification of important subnetworks in two-mode networks with respect to selected node properties.

We combine the ideas from generalized cores in one-mode networks and from (p, q)-cores for two-mode networks into the notion of *generalized two-mode cores*. We developed and implemented an algorithm for identifying generalized two-mode cores for selected node properties and given thresholds for both subsets of nodes. We also propose an algorithm to find the nested generalized two-mode cores for one fixed threshold value.

In the next section we survey the works that contain the ideas we used for the development of our method. In Section 3 we present an algorithm for identifying the generalized two-mode cores. We list some node properties that are used as measures of importance. We also present some properties of generalized two-mode cores. We prove that for equivalent properties measured in ordinal scales the sets of generalized two-mode cores are the same. The algorithm, the proof of its correctness, and a simple analysis of its complexity are presented in Section 4. In Section 5 some results obtained in analyses of real-life data are presented.

2. Related work

The notion of k-core was introduced by Seidman (1983). Let $\mathcal{G} = (\mathcal{V}, \mathcal{L})$ be a graph with $n = |\mathcal{V}|$ nodes and $m = |\mathcal{L}|$ links. Let k be a fixed integer and let $deg(v)$ be the degree of a node $v \in \mathcal{V}$. A subgraph $\mathcal{H}_k = (\mathcal{C}_k, \mathcal{L}|\mathcal{C}_k)$ induced by the subset $\mathcal{C}_k \subseteq \mathcal{V}$ is called a *k-core* iff $deg_{\mathcal{H}_k}(v) \geq k$, for all $v \in \mathcal{C}_k$, and $\mathcal{H}_k$ is the maximal such subgraph. If we replace the degree with some other node property, we get the notion of *generalized cores* as it was introduced in Batagelj and Zaveršnik (2010). The node property can be a node degree, maximum of incident link weights, sum of incident link weights, etc. They are described in more details in Section 3.

The other possible generalization of k-cores is their extension to two-mode networks. The notion of (p, q)-cores was introduced in Ahmed et al. (2007). A subset $\mathcal{C} \subseteq \mathcal{V}$ determines a (p, q)-core in a

* Corresponding author. Tel.: +386 31421590.
E-mail addresses: monika@hruska.si (M. Cerinšek), vladimir.batagelj@fmf.uni-lj.si (V. Batagelj).
URL: http://vladowiki.fmf.uni-lj.si/ (V. Batagelj).

http://dx.doi.org/10.1016/j.socnet.2015.04.001
0378-8733/

Table 1
Examples of property functions

Formula	Description
$f_1(v,\mathcal{C}) = deg_{\mathcal{C}}(v)$	Degree of a node v within $\mathcal{C}$
$f_2(v,\mathcal{C}) = \mathrm{indeg}_{\mathcal{C}}(v)$	Input degree of a node v within $\mathcal{C}$
$f_3(v,\mathcal{C}) = \mathrm{outdeg}_{\mathcal{C}}(v)$	Output degree of a node v within $\mathcal{C}$
$f_4(v,\mathcal{C}) = \mathrm{indeg}_{\mathcal{C}}(v) + \mathrm{outdeg}_{\mathcal{C}}(v)$	For a directed network $f_4 = f_1$
$f_5(v,\mathcal{C}) = \mathrm{wdeg}_{\mathcal{C}}(v) = \sum_{u\in N(v,\mathcal{C})} w(v,u)$	Sum of weights of links within $\mathcal{C}$ that have a node v for an end node.
$f_6(v,\mathcal{C}) = \mathrm{mweight}_{\mathcal{C}}(v) = \max_{u\in N(v,\mathcal{C})} w(v,u)$	Maximum of weights of all links within $\mathcal{C}$ that have a node v for an end node.
$f_7(v,\mathcal{C}) = \mathrm{pdeg}_{\mathcal{C}}(v) = \frac{deg_{\mathcal{C}}(v)}{deg(v)}$ if $deg(v) > 0$ else $f_7(v,\mathcal{C}) = 0$	Proportion of $N(v,\mathcal{C})$ in $N(v)$
$f_8(v,\mathcal{C}) = \mathrm{density}_{\mathcal{C}}(v) = \frac{deg_{\mathcal{C}}(v)}{\max_{u\in N(v)} deg(u)}$ if $deg(v) > 0$ else $f_8(v,\mathcal{C}) = 0$	Relative density of the neighborhood of a node v within $\mathcal{C}$
$f_9(v,\mathcal{C}) = \mathrm{degrange}_{\mathcal{C}}(v) = \max_{u\in N(v,\mathcal{C})} deg(u) - \min_{u\in N(v,\mathcal{C})} deg(u)$	Range of degrees of neighbors of a node v within $\mathcal{C}$ with respect to their degrees.
$f_{10}(v,\mathcal{C}) = \mathrm{tdegrange}_{\mathcal{C}}(v) = \max_{u\in N^+(v,\mathcal{C})} deg(u) - \min_{u\in N^+(v,\mathcal{C})} deg(u)$	Total range of degrees of neighbors of a node v
$f_{11}(v,\mathcal{C}) = \mathrm{pweight}_{\mathcal{C}}(v) = \frac{\sum_{u\in N(v,\mathcal{C})} w(v,u)}{\sum_{u\in N(v)} w(v,u)}$ if $\sum_{u\in N(v)} w(v,u) > 0$ else $f_{11}(v,\mathcal{C}) = 0$	Proportion of weights of links with a node v as an end node that have the other end node within $\mathcal{C}$
$f_{12}(v,\mathcal{C}) = \mathrm{triangles}_{\mathcal{C}}(v) =$ Number of triangles through a node v with all three nodes in $\mathcal{C}$	
$f_{13}(v,\mathcal{C}) = \mathrm{sum}\,\mathcal{C}(v,t) = \sum_{u\in N(v,\mathcal{C})} t(u)$	
$f_{14}(v,\mathcal{C}) = \max\mathcal{C}(v,t) = \max_{u\in N(v,\mathcal{C})} t(u)$	

two-mode network $\mathcal{N} = ((\mathcal{V}_1,\mathcal{V}_2),\mathcal{L})$, $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$ iff in the subnetwork $\mathcal{K} = ((\mathcal{C}_1,\mathcal{C}_2),\mathcal{L}|\mathcal{C})$, $\mathcal{C}_1 = \mathcal{C}\cap\mathcal{V}_1$, $\mathcal{C}_2 = \mathcal{C}\cap\mathcal{V}_2$ induced by $\mathcal{C}$ it holds that for all $v \in \mathcal{C}_1 : deg_{\mathcal{K}}(v) \geq p$ and for all $v \in \mathcal{C}_2 : deg_{\mathcal{K}}(v) \geq q$, and $\mathcal{C}$ is the maximal such subset in $\mathcal{V}$.

We combined the ideas from generalized cores and (p, q)-cores into the notion of generalized two-mode cores. Generalized two-mode cores are defined similarly to (p, q)-cores with one exception – instead of using the degree of nodes, we now allow also some other properties of nodes. The properties of nodes on subsets $\mathcal{V}_1$ and $\mathcal{V}_2$ can be different. The detailed definition is given in Section 3.1.

Other types of two-mode subnetworks were discussed in the literature. A bipartite core is defined as a complete two-mode subnetwork (Xu et al., 2011). Bipartite cores are determined by the size of each subset of vertices. Similar methods are varieties of a community detection in two-mode networks: with maximization of monotone function (Sozio and Gionis, 2010), with a dual projection (Melamed, 2014), with properties of the eigenspectrum of the network's matrix (Barber, 2007), with label propagation and recursive division of the two types of nodes (Xin, 2009), with the stochastic block modeling (Larremore et al., 2014), and many others. Another very similar method is a bipartite clustering (Baier and Wernecke, 2003; Van Mechelen et al., 2004). The substantial difference is that these methods are determining a clustering of the whole set of nodes and our method determines only an important subset.

The generalized two-mode cores depend on selected node properties that are expressing different aspects of the network structure (for example, the intensity of links). They are also using different criteria. Therefore our method represents a new approach to two-mode network analysis. It does not represent an improvement of any existing method, but a generalization of (p, q)-cores.

3. Algorithms for generalized two-mode cores

As mentioned in Section 1 the algorithms for identifying k-cores, generalized cores, and (p, q)-cores have already been developed (Ahmed et al., 2007; Batagelj and Zaveršnik, 2010; Seidman, 1983). We propose a new algorithm, which combines and extends the ideas from generalized cores in one-mode networks and from (p, q)-cores in two-mode networks. Besides implementing the new algorithm, we also prove its correctness and analyze its complexity. For testing the usefulness of the method we applied it on real networks.

3.1. Properties of nodes

For a network $\mathcal{N} = (\mathcal{V},\mathcal{L},w)$ and a weight function $w : \mathcal{L} \to \mathbb{R}^+$ a *property function* $f(v,\mathcal{C}) \in \mathbb{R}_0^+$ is defined for all $v \in \mathcal{V}$ and $\mathcal{C} \subseteq \mathcal{V}$. A subset $\mathcal{C}$ induces the subnetwork to which the evaluation of the property function is limited. In an undirected network it holds: $w(u,v) = w(v,u)$ for all pairs of nodes $u, v \in \mathcal{V}$.

Let us denote the neighborhood of a node v as $N(v)$ and the neighborhood of a node v within the subset $\mathcal{C}$ as $N(v,\mathcal{C}) = N(v)\cap\mathcal{C}$. The neighborhood of a node v within the subset $\mathcal{C}$ including v we denote as $N^+(v,\mathcal{C}) = N(v,\mathcal{C})\cup\{v\}$. Let us also denote a measurement on nodes (degree, centrality, etc.) as $t : \mathcal{V} \to \mathbb{R}_0^+$.

We say that the property function $f(v,\mathcal{C})$ is *local* iff

$$f(v,\mathcal{C}) = f(v,N(v,\mathcal{C})) \quad \text{for all } v\in\mathcal{V} \text{ and } \mathcal{C}\subseteq\mathcal{V}.$$

The property function $f(v,\mathcal{C})$ is *monotonic* iff

$$\mathcal{C}_1 \subset \mathcal{C}_2 \Rightarrow \forall v\in\mathcal{V} : f(v,\mathcal{C}_1) \leq f(v,\mathcal{C}_2).$$

Some node properties $(f_1 - f_{10})$ were proposed in Batagelj and Zaveršnik (2010). In Table 1 are listed examples of property functions.

All the listed functions have the property $f(v,\emptyset) = 0$ for all $v \in \mathcal{V}$.

It can easily be verified that all the listed property functions are local and monotonic. An example of a non-monotonic function would be the average weight

$$f(v,\mathcal{C}) = \frac{1}{deg_{\mathcal{C}}(v)} \sum_{u\in N(v,\mathcal{C})} w(v,u)$$

for $N(v,\mathcal{C}) \neq \emptyset$, otherwise $f(v,\mathcal{C}) = 0$. An example of a non-local function is the number of cycles or closed walks of length k, $k \geq 4$, through a node.

3.2. Generalized two-mode cores

Definition 3.1. Let $\mathcal{N} = ((\mathcal{V}_1,\mathcal{V}_2),\mathcal{L},(f,g),w)$, $\mathcal{V} = \mathcal{V}_1\cup\mathcal{V}_2$ be a finite two-mode network – the sets $\mathcal{V}$ and $\mathcal{L}$ are finite. Let $\mathcal{P}(\mathcal{V})$ be a power set of the set $\mathcal{V}$. Let functions f and g be defined on the network $\mathcal{N}$: $f, g : \mathcal{V}\times\mathcal{P}(\mathcal{V}) \longrightarrow \mathbb{R}_0^+$.