Interdisciplinarity of Slovenian research

Borut Lužar

joint work with Zoran Levnajić, Matjaž Perc and Janez Povh

Faculty of Information Science, Novo mesto
&
Institute of Mathematics, Physics and Mechanics, Ljubljana,
Slovenia

November 7, 2013

Motivation

Question: How do the scientists from different science areas connect?

One of many answers: Detect communities of Slovenian researchers and evaluate their interdisciplinarity in different time periods.

Data

- Sicris & Cobiss data on Slovenian researchers and their publications
- Networks for 8 time periods: 1960-1965 ($\mathcal{G}_{65}$), 1960-1970 ($\mathcal{G}_{70}$),..., 1960-2000 ($\mathcal{G}_{00}$) (created using the application designed in the framework of the project *Co-authorship networks of slovenian scholars*)
- The vertex set of $\mathcal{G}_t$ are the researchers that published a paper or a monograph in the given period
- Two researchers in $\mathcal{G}_t$ are connected if they coauthored a publication (we ignore the number of common publications)

Communities

- No strict definition. A subgraph of a graph which is more dense might be considered as a community within a network.
- Our networks are not connected. Every component with at least 5 vertices is a community. The largest component is divided further.
- For the separation of the largest component we use the algorithm by Girvan and Newman using the edge-betweenness approach.

Algorithm

- The *edge-betweenness* of an edge is the number of shortest paths between the pairs of vertices in the graph traversing the edge.
- The algorithm proceeds by repetitively removing an edge with the maximum edge-betweenness from the current graph until some stopping criteria is reached. Each component of the final graph is a community.
- The stopping criteria is problematic. We decided to stop the execution when the maximum edge-betweenness of a graph is greater than $n - 1$, where n is the number of vertices in the initial graph - the largest component of $\mathcal{G}_t$

Data

A set of (non-overlapping) communities $\mathcal{C}_t$ which we analyze is obtained. For the analysis also classification data is needed. Every registered researcher is assigned a science area he is involved in. Sicris uses a classification into 7 sciences:

1. Natural sciences and mathematics
2. Engineering sciences and technologies
3. Medical sciences
4. Biotechnical sciences
5. Social sciences
6. Humanities
7. Interdisciplinary studies (just a wish)

To every community C a seven component vector $S(C)$ is assigned - the i-th component represents a relative number of researchers in the community with the i-th science assigned.

Interdisciplinarity measures

To measure interdisciplinarities of the communities, we introduce three measures:

1. single-bound interdisciplinarity;
2. ordinary interdisciplinarity;
3. normalized interdisciplinarity.

Single-bound interdisciplinarity - $B(\mathcal{G}_t)$

A simple measure, defined as

$$B(\mathcal{G}_t) = \frac{\#\text{ interdisciplinary communities}}{|\mathcal{C}_t|},$$

where by an *interdisciplinary community* we mean every community C with the largest component of $S(C)$ smaller than some given bound - we chose 0.8.

Ordinary interdisciplinarity - $O(\mathcal{G}_t)$

For each community C we compute its $O(C)$ as:

$$O(C) = -\frac{7}{6}M(C) + \frac{7}{6}\,.$$

Notice that $O(C) = 0$ if $M(C) = 1$ and $O(C) = 1$ if $M(C) = \frac{1}{7}$, which is the smallest possible value for $M(C)$. The ordinary interdisciplinarity of $\mathcal{G}_t$ is defined as

$$O(\mathcal{G}_t) = \frac{1}{|\mathcal{C}_t|}\sum_{C\in\mathcal{C}_t} O(C)\,.$$

Normalized interdisciplinarity - $N(\mathcal{G}_t)$

Normalized interdisciplinarity $N(C)$ of a community C is defined as

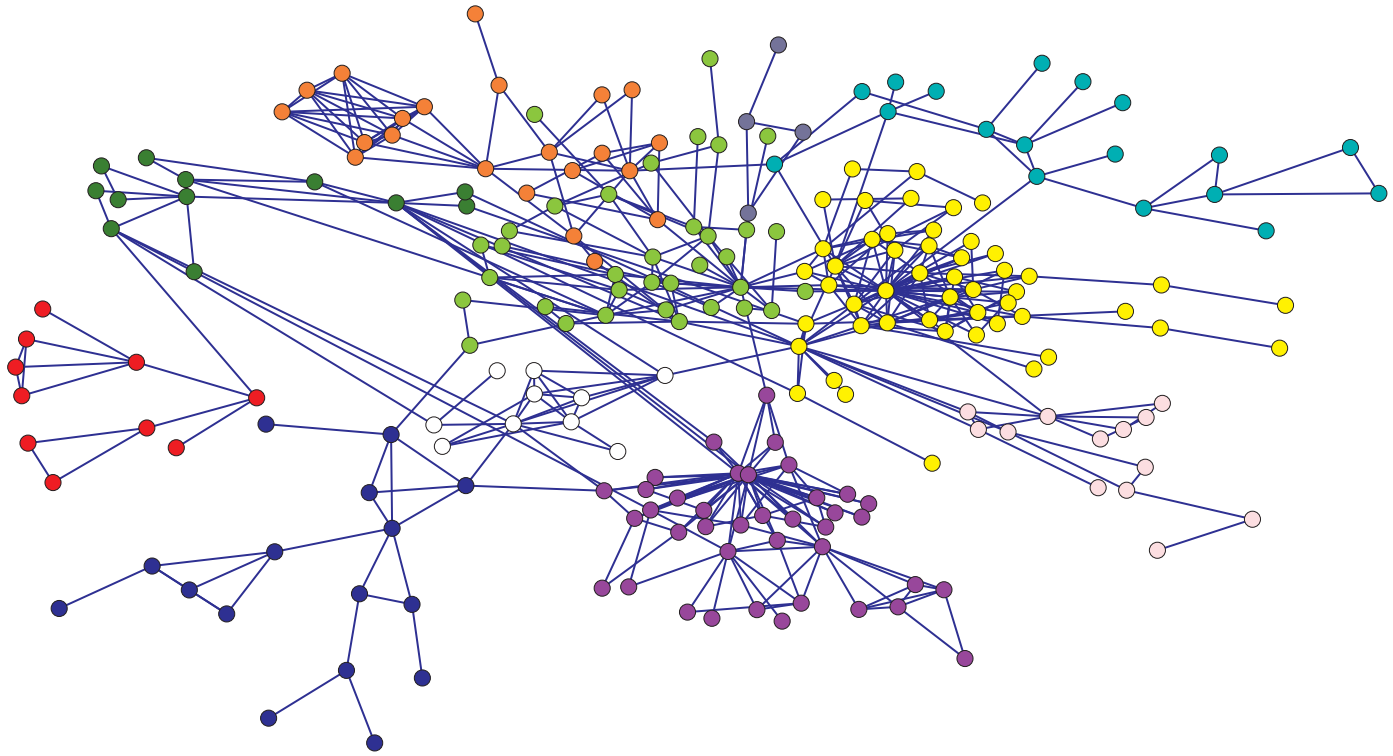
$$N(C) = O(C) \cdot |C| \,.$$

The normalized interdisciplinarity of $\mathcal{G}_t$ is then computed as

$$N(\mathcal{G}_t) = \frac{1}{v(\mathcal{C}_t)} \sum_{C \in \mathcal{C}_t} N(C) \,.$$

Results

Separation of the largest component of $\mathcal{G}_{75}$ into communities:



Results

network	n	m	l	c	n_c
$\mathcal{G}_{65}$	194	70	18	2	12
$\mathcal{G}_{70}$	454	321	76	12	144
$\mathcal{G}_{75}$	920	947	229	29	432
$\mathcal{G}_{80}$	1586	2243	690	41	913
$\mathcal{G}_{85}$	2242	3739	1195	45	1397
$\mathcal{G}_{90}$	3799	7838	2408	60	2544
$\mathcal{G}_{95}$	5689	13580	3971	74	4076
$\mathcal{G}_{00}$	7925	22600	5888	91	6000

Table: Basic properties of analyzed networks. Here n, m, l, c, and n_c are the number of vertices in the network, the number of edges in the network, the size of the largest component, the number of analyzed communities and the number of vertices in the analyzed communities, respectively.

Results

network	$B(\mathcal{G}_t)$	$O(\mathcal{G}_t)$	$N(\mathcal{G}_t)$
$\mathcal{G}_{65}$	0.000	0.082	0.095
$\mathcal{G}_{70}$	0.357	0.250	0.244
$\mathcal{G}_{75}$	0.760	0.327	0.341
$\mathcal{G}_{80}$	0.634	0.312	0.355
$\mathcal{G}_{85}$	0.689	0.338	0.365
$\mathcal{G}_{90}$	0.683	0.337	0.365
$\mathcal{G}_{95}$	0.622	0.330	0.332
$\mathcal{G}_{00}$	0.560	0.300	0.359

Table: The values of the three measures of interdisciplinarity for every network.

Thank you!