Properties of Communication Graph of Academic Web

A.A. Pechnikov [1][0000-0002-0683-0019]

[1] Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences, 11, Pushkinskaya str., 185910, Petrozavodsk, Russia
pechnikov@krc.karelia.ru

Abstract. Web graph is the most popular model of real web fragments used in Web science. The study of communities in the web graph contributes to a better understanding of the organization of the web fragment and the processes taking place in it. Communities of the web-graph fragments of the real Web are often poorly differentiated and more difficult meaningful interpretation. It is proposed to select in the web graph a communication graph containing only those vertices (and arcs between them) that have counter arcs, and already in it to investigate the problem of splitting into communities. By analogy with social studies, the ties implemented through the edges in the communication graph are proposed to be called "strong" and all others "weak". Thematic communities with meaningful interpretation are built on strong ties. At the same time, weak ties facilitate communication between sites that do not have common features in the sphere of activity, geography, subordination, etc., and basically keep the web fragments connected even in the absence of strong ties. Experiments carried out for a fragment of the academic Web of Russia show the possibility of meaningful interpretation of the results and the prospects of this approach.

Keywords: Web graph, Communication graph, Community in graph, Strength of ties.

1 Introduction

The study of graphs of the real (and virtual) world is an important task in many fields, such as biology, sociology, social networks, webometrics, and many others, because they allow you to understand the structure of objects and analyze their properties. It has long been known that University and academic fragments of the Web both in Russia and other countries [1–3] have quite specific properties characterizing their structure. In particular, the corresponding web graph has a large strongly connected component (SCC) and a significant number of "hanging" sites (having either only links made from them, or, more rarely, links made only to them).

SCC itself is an interesting object of research, allowing to establish, for example, the presence or absence of the property of the "small world" [4], which, however, does little to explain the processes of hyperlinks between sites fragments of the Web. In this sense, much more food for thought is provided by structural elements of the graph, such as site communities, when more links are made "within" communities than between communities. But again, attempts to partition the vertices that make up the maximum connectivity component of a web graph into communities lead to difficult to interpret results. The main idea considered in this work is to build a "hard framework" for a fragment of the Web, leaving only those sites that have counter hyperlinks, and already on this "framework" to check the properties of the partition into communities (such a framework will be called a communication graph). And already further by means of known algorithms the question of structure of communities of tops, the communication graph and the "weak" web graph from which the communication graph is removed is investigated. As a real object of research, the scientific and educational fragment of the Web is taken, for which a meaningful interpretation of the results is given.

2 Basic Concepts, Methods and Tools

The target set of sites is specified by a direct enumeration of their domain names, and as a result of scanning, all the hyperlinks linking them are found. In the case where sites belong to one type of activity, such a set is called thematic. Accordingly, a (thematic) fragment of the Web is a target set of sites and a set of hyperlinks linking them [3].

A web graph of Web fragment is a directed graph without loops and multiple arcs, whose vertex set is represented by the target site set, and whose arc set is constructed as follows: two vertices are connected by an arc if there is at least one hyperlink linking the corresponding sites.

A community (cluster, module, group) of a graph can be informally defined as a set of vertices with more arcs between them than with the other vertices of the graph [5].

More strictly, the partition of a graph into communities can be defined through the notion of modularity. Modularity is a property of a graph and some subdivision of it into subgraphs (community modules). The modularity measure shows how qualitative a given partition is in the sense that there are many arcs lying inside subgraphs-communities and few arcs lying outside subgraphs (connecting communities together).

The modularity measure Q can be given by the following formula [5]:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right) * \delta(c_i, c_j).$$

Here m is the number of arcs in the graph; A_{ij} is the element of the adjacency matrix of a graph; k_i, k_j is the degree of the respective nodes; $\delta(c_i, c_j)$ is Kronecker's symbol calculated by the formula:

$$\delta(c_i, c_j) = \begin{cases} 1: if\ c_i = c_j \\ 0: if\ c_i \neq c_j \end{cases} \quad ,$$

where c_i, c_j are the class labels of the corresponding vertices.

Accordingly, the best partition into communities is considered to be the partition on which the maximum Q is achieved.

BeeCrawler [6] and external hyperlink databases [7] were used for data collection and analysis. The open-source graph visualization platform Gephi [8] was used for the study of web graphs, which, among many, implemented programs for calculating the modularity coefficient and searching for communities in the graph using the algorithms proposed in [9].

3 Fragment of the Scientific and Educational Web of Russia

The initial target set of sites of a fragment of the scientific and educational Web of Russia dates from the end of 2017, that is, after the Russian Academy of Medical Sciences and the Russian Academy of Agricultural Sciences (RAAS) joined the Russian Academy of Sciences (RAS), but in the process of creating enlarged structures such as federal research centers, as well as enlarging universities.

Such a “footprint” allows us to expect on the fact that a fragment of the Web contains fairly stable links between sites that have been formed over the past few years. The initial target set contains 867 objects (it is websites of 279 universities and 588 research institutions). Scientific institutions are understood as organizations of the RAS from the level of institutes to regional scientific centers and boards of the RAS. Website of the RAS itself (www.ras.ru) in the target set is not included, because it is a powerful communicator, significantly distorting the common picture. Further, for the sake of brevity, we will call scientific institutions "institutes".

Scanning all 867 sites allows you to get about 20,000 hyperlinks made between these sites, including multiple links. Replace multiple links to single arcs and get a web graph containing 867 vertices and 5030 arcs. A test for connectivity and strong connectivity shows the presence of 73 isolated vertices and 236 hanging (36 vertices do not have incoming arcs and 200 (!) - outgoing). In terms of content, we note that 73 isolated vertices in the vast majority relate to the sites of institutes that were previously part of the RAAS.

The only SCC contains 534 vertices and 4026 arcs. A web graph equal to the maximum SCC has a diameter equal to 10 and a modularity coefficient of 0.398 [10], and is divided into 7 communities containing 40 to 139 vertices. The modularity coefficient indicates a low tendency of sites to organize into communities: the values of Q belong to the segment [-1, 1], and clustering is considered "good" somewhere when Q is greater than 0.6.

However, one of the communities built has a good meaningful explanation. It includes 40 vertices corresponding to 4 university sites and 36 institute sites, and a total of 97 arcs (which on average is significantly less than in the SCC web graph). Of the 40 websites, 35 belong to the current Department of agricultural sciences RAS, as well as to the Krasnoyarsk scientific center of Siberian Branch of RAS, Kemerovo Institute of technology of food industry, Kuban state technological University, Voronezh state university of forestry and St. Petersburg Mining University.

Give comments:

* Krasnoyarsk scientific center for the period of the study managed to become a Federal research center, which includes 2 agricultural institutes;

* Kemerovo Institute of technology of food industry is close to agriculture;

* Kuban state technological university has close ties with agricultural institutes of Kuban;

* Voronezh state forestry university has a hyperlink on its website to the website of the Central scientific agricultural library (which falls under the term "institutes»);

* St. Petersburg mining university has on its website a hyperlink to the website of the All-Russian Institute of plant genetic resources named after N.I. Vavilov, where there is a section of the journal "Agro-industrial complex", issued by the University library.

Apart from the last two cases, we can say that this community has a pronounced agro-industrial theme.

4 The Communication Graph of the Web Graph

Newman and coauthors, when solving the problem of splitting a graph into communities, avoid graph orientation when dealing with web graphs.

Quoting from [10, p. 026113-5]: "…Some networks are directed, i.e., their edges run in one direction only. The world wide web is an example; links in the web point in one direction only from one web page to another. … However, we have found that in many cases it is better to ignore the directed nature of a network in calculating community structure. Often an edge acts simply as an indication of a connection between two nodes, and its direction is unimportant…". And if direction (orientation) matters? - Let's look at fig. 1, which shows agro-industrial community.

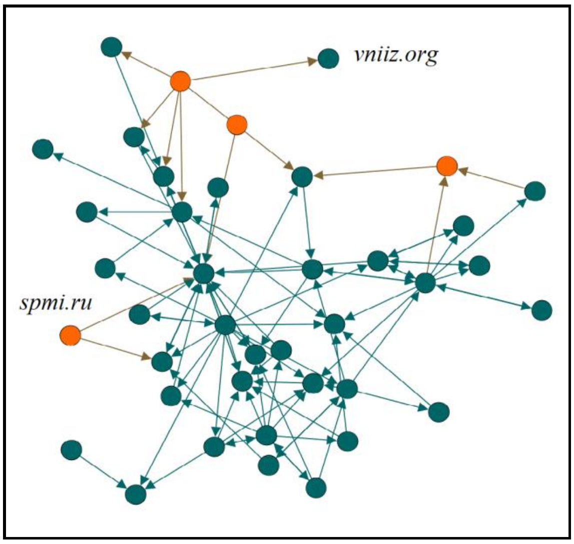

Fig. 1. Web graph "agro-industrial community".

There are questions. Can the website *vniiz.org* (All-Russian research institute of grain and products of its processing) consider a full member of the community when it does not have hyperlinks to other sites? Can the website *spmi.ru* (St. Petersburg mining university) consider a member of a community when there are no links to it from other members?

The term "community" (*Gemeinschaft*) originated in German sociology at the end of the XIX century and implies the joint activity of its participants with common goals [11], and in this context the answer to the formulated questions should be negative. Therefore, we will further consider an undirected graph, which by construction implies “joint activity”, namely, counter hyperlinks.

A communication graph of a web graph is an undirected graph that has the same set of vertices as the web graph, and the set of edges is constructed according to the following rule: an edge (i, j) belongs to the set of edges of a communication graph if and only if in the web graph there exist arcs (i, j) and (j, i). A communication graph constructed in this way can have several connected components and / or isolated vertices. In this case, we exclude isolated vertices (since they do not affect connectivity), and then study the connected components individually, starting with the maximum.

The communication graph of the web graph of the scientific and educational Web contains 313 vertices and 468 edges (Fig. 2).

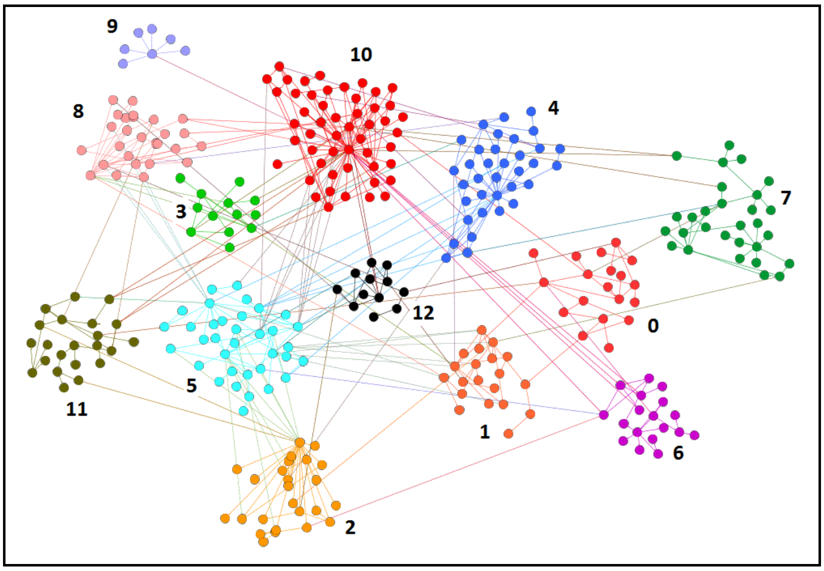

Fig. 2. Communication graph of scientific and educational Web.

Of the 313 vertices, 67 belong to universities and 246 to institutes, i.e. the proportion of universities in the communication graph has decreased by one tenth. The coloring of the vertices is given according to the resulting partition into communities.

The modularity coefficient of this partition is 0.695, which is quite high. Built 13 communities contain from 7 to 51 vertices, of which 4 can be accurately identified by one of two features: geography or scientific direction.

Some of the built communities can be given a meaningful (thematic) interpretation.

Two "geographical" communities are clearly distinguished:

No. 2 – "Ural": 1 university (Perm Polytechnical), 23 institutes of the Ural Branch of RAS and the Research center of fiber optics of RAS from Moscow;

No. 10 – "Siberia": 4 Siberian universities, 47 Siberian institutes and 1 Institute from Sevastopol.

"Humanitarian" community No. 12 contains 10 institutes (archaeology, ethnography, linguistic research, linguistics, history, world literature, Slavic studies, philosophy) and two universities (Altai State University and Gorno-Altaisk State University).

"Medical" community No. 9 is a "star" of websites of medical institutes of Siberia and the Far East, linked to the website of the Siberian Department of medical Sciences, but not related to each other.

Members of any communication graph community are present in the web graph by construction. However, it is not necessary that all members from the same communication graph community belong to the same web graph community. The reverse is not true: the sites participating in the agro-industrial community of the web graph are completely absent from the communication graph.

Now remove all pairs of counter arcs corresponding to the edges of the communication graph from the SCC web graph. The resulting "weak" web graph is large enough – 528 vertices and 3090 arcs, but has a low modularity coefficient of 0.356 and is broken up by 8 communities that do not have a convincing meaningful interpretation.

Find the maximum SCC of the "weak" web graph, remove all vertices not included in the SCC and get the SCC "weak" web graph with 461 vertices and 2744 arcs. Again, the modularity coefficient remains low.

5 Discussion of Results. Strong and Weak Ties

More than 45 years ago, Granovetter [12] proposed the concept of the strength of social ties. All connections are divided into two categories, strong and weak, in order to formalize interpersonal relationships based on the duration and frequency of contacts. For example, a strong connection is inherent in friends, and a weak one is inherent in neighbors. Let's try to use the concept of communication strength as an analogy for sites, although it is initially clear that the analogy is far from complete, since Granovetter believes that the connections are symmetrical.

Take a pair of vertices *i*, *j* of a directed graph. If there are arcs (*i*, *j*) and (*j*, *i*) in the graph, then we will talk about a strong connection of vertices *i*, *j*.

If for vertices *i*, *j* there is only one of the arcs (*i*, *j*) or (*j*, *i*), then we will talk about a weak connection between vertices *i* and *j*.

Let's look at the dynamics of changes in the proportion of institutes and universities in the considered graphs, for which we will summarize the data in Table 1.

Table 1. Proportion of sites of institutes and universities in graphs.

Graph	Number of vertices	Number of arcs (edges)	Proportion of institutes	Proportion of universities	Diameter	Average path length
Initial web graph	867	5030	0,68	0,32	-	-
SCC web graph	534	4026	0,67	0,33	10	3,6
Communication web graph	313	936 (468)	0,78	0,22	9	3,45
"Weak" web graph	528	3090	0,66	0,34	-	-
SCC "weak" web graph	461	2744	0,63	0,37	11	4,2

The initial scientific and educational web graph like the SCC web graph contain strong and weak arcs, the communication graph lacks weak arcs, and the" weak "web graph and the SCC "weak" web graph lack strong arcs.

From Table 1 it can be seen that the institutes/universities ratio is practically the same for all graphs having weak or strong and weak arcs, but changes dramatically for a graph containing only strong arcs. It can be concluded that strong ties are more inherent in institutions than universities.

In [12, p. 1362] it is noted that " ... the stronger the ties the more similar individuals are to each other in different aspects". This is confirmed in the case of the communication graph of the scientific and educational Web. We can say that strong connections contribute to the emergence of sustainable thematically "similar" communities. Although there are few such communities, only 4 out of 13, they have a clear meaningful interpretation.

However, if the communication graph is "built up" to the SCC web graph (and even more so, to the initial web graph), it turns out that none of the four communities of the communication graph not only was not the basis for communities in web graphs, but even their members were in different communities. It seems that the weak links lead to the erosion of communities.

Recall that in the SCC web graph we found an " agro-industrial" community, and in the initial web graph we found 73 isolated vertices also related to agricultural institutes. By linking these two results, the following explanation for this exception can be offered. The websites of the former RAAS simply did not have time to establish links with other sites of the target set. Therefore, those RAAS sites that were somehow connected to each other had few links "outside" and organized a community, and the rest remained isolated.

What is the role of weak links? According to Granovetter, weak connections can be characterized as a system of irregular contacts that do not cover the friends of an individual, but connect him with members of other closely related groups in which he does not belong. From this it follows that an individual can establish relationships with people from a large number of mutually disjoint groups, while not being a member of each of them.

In this sense, weak connections play the role of "bridges" in the graph. Recall that a bridge is an edge of a (undirected) graph whose removal increases the number of connectivity components [13]. Granovetter in [12] says that strong ties are almost never bridges, and as a rule, weak bonds are bridges. In our case, indirect confirmation of this fact is the diameter and average path length, which are almost the same in the SCC "weak" web graph, SCC web graph and communication graph (Table 1). That is, the removal of strong ties has almost no effect on the diameter of the graphs, which means that the arcs corresponding to them are often not bridges.

Hence the conclusion that the weak ties of the fragment of the Web serve to establish contacts of sites from disjoint groups, which is very important in the sense of obtaining new information that does not lie in the circle of interests of this group.

The work was supported by The Russian Foundation for Basic Research, project 18-07-00628-a.

References

1. Thelwall, M., Wilkinson, D.: Graph structure in three national academic Webs: Power laws with anomalies. Journal of the American Society for Information Science and Technology 54(8), pp. 706–712 (2003).
2. Ortega, J.L., Aguillo, I.F.: Visualization of the Nordic academic web: Link analysis using social network tools. Information Processing and Management 44(4), pp. 1624–1633 (2008).
3. Pechnikov, A.A.: Metody issledovanija reglamentiruemyh tematicheskih fragmentov Web. Trudy Instituta sistemnogo analiza Rossiiskoi akademii nauk. Serija: Prikladnye problem upravlenija makrosistemami 59, pp. 134–145 (2010).
4. Watts, D.J.; Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393, pp. 440–442 (1998).
5. Ermolin, N.A., Mazalov, V.V., Pechnikov, A.A.: Teoretiko-igrovye metody nahojdenija soobschestv v akademicheskom Webe. Trudy SPIIRAN 55, pp. 237–254 (2017).
6. Pechnikov, A.A., Chernobrovkin, D.I. Adaptive Crawler for External Hyperlinks Search and Acquisition. Automation and Remote Control 75(3), pp. 587–593 (2014).
7. Golovin, A.S., Pechnikov, A.A.: Baza dannyh vneshnih giperssylok dlja issledovanija fragmentov Weba. In: Informacionnaja sreda vuza XXI veka: materialy VII Vserossiiskoi nauchno-prakticheskoi konferencii, pp. 55–57. PetrSU, Petrozavodsk (2013).
8. The Open Graph Viz Platform, https://gephi.org, last accessed 2019/12/03.
9. Blondel, V.D., Guillaume, J-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, P10008 (2008).
10. Newman, M.E., Girvan, M. Finding and evaluating community structure in networks. Physical Review E. 69(2), P026113 (2004).
11. Levkina, L.I.: Social'no-istoricheskaja rol' soobschestv. Rusains, Moscow (2016).
12. Granovetter, M.S.: The Strength of Weak Ties. The American Journal of Sociology 78(6), pp. 1360–1380 (1973).
13. Harary, F.: Graph Theory. Addison-Wesley, Boston (1969).