Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 31 (2014) 892 – 899

Procedia Computer Science

Information Technology and Quantitative Management (ITQM 2014)

Connectivity analysis of computer science centers based on scientific publications data for major Russian cities

Fedor Krasnov[a], Evgeniya Vlasova[b], Rostislav Yavorskiy[b]*

[a]Skolkovo Foundation, Krasnopresnenskaya nab. 12-9, floor 25, Moscow 143026, Russia
[b]Higher School of Economics, Myasnitskaya 20, Moscow 101000, Russia

Abstract

In this paper we study connectivity between leading universities and academic institutions in the area of computer science for Moscow and Saint Petersburg. The research is based on scientific publications data available from http://eLibrary.ru. The focus is restricted to scientific papers in IT related areas, namely Informatics and Cybernetics, published during 2011-2013. We assume that two organizations are connected via co-authorship, *a*-linked, if there exists at least one paper, which has employees from these organizations among co-authors. In order to detect closely related yet not collaborating organizations formal context *organizations-publishers* is studied. We say that two institutions are connected via publishers, *p*-linked, if the pair forms extent of a formal concept.

Selection and peer-review under responsibility of the Organizing Committee of ITQM 2014.

Keywords: urban scientific communities, community measurements, connectivity of a community

1. Introduction

Our general goal is to develop collection of efficient and easy-to-use methods and tools for measuring key characteristics of urban professional communities, such as its strength, vividness, maturity level, connectivity, size, major clusters, internal structure etc.

This paper continues our research on urban IT communities we started in [1-3]. It is well known that strong computer science centers play crucial role in development of IT industry by feeding it with ideas, new technologies and talented people.

* Corresponding author. Tel.: +7-964-783-2485; fax: +7-495-628-79-31.
E-mail address: ryavorsky@hse.ru.

Selection and peer-review under responsibility of the Organizing Committee of ITQM 2014.
doi:10.1016/j.procs.2014.05.341

In [3] the structure of research communities for 12 regional centers with population of more than 1 million people was considered. In this paper we analyze the inner community structure of Moscow and Saint Petersburg.

1.1. Study of innovative cities

Innovative cities became a hot topic in the last decades of XX century when different countries attempted to replicate the success of the Silicon Valley on their territory, see [4-5].

1.2. Scientific Communities of Practice

Different studies show that interconnected and active communities create a fruitful environment for collaboration, see [6-13] for more.

1.3. Research goals and the structure of this paper

In this paper we consider two types of connections between institutions, *a*-links and *p*-links. The first, *a*-link, is based on the co-authorship relation. The second, *p*-link, indicates that two organizations are strongly connected via the publishers. Formal definitions are given in section 2. To model connectivity we use a model based on Formal Concept Analysis (FCA).

2. Formal definitions

In this section the organizations connectivity graph is defined for a given city.

2.1. Nodes of the organizations connectivity graph

Let D stand for a dataset of scientific publications. For a given city S we denote by V the set of all organizations v located in S such that D contains at least one paper authored or co-authored by a person working in v.

2.2. Co-authorship of scientific publications, a-links

We say that two organizations v and w from V are connected via co-authorship, *a-linked*, if there exists at least one paper in D, which has employees from v and w among co-authors. The number of such publications in D defines *strength* of the corresponding *a*-link.

2.3. Connectivity via publishers, p-links

We assume that dataset D contains publisher information for all papers. Consider a formal context (V, P, I), where V is the set of institutions defined above, P is the set of publishers, I is the binary relation between institutions and publishes, $(v, p) \in I$ indicates that there exists at least one paper in D co-authored by person from v, which is published by p.

Now, for this context we can compute all formal concepts with binary extent, that is all pairs (V_i,P_i) such that:

- $V_i \subseteq V$, $P_i \subseteq P$, $i \in \{1,...,n\}$;

- $V_i' = P_i$, $P_i' = V_i$, where prime stands for the Galois operator;
- $|V_i| = 2$ for all $i \in \{1,...,n\}$.

Finally, we say that two institutions *v* and *w* are connected via publishers, *p-linked*, if $\{v, w\} = V_i$ for some *i* in $\{1,...,n\}$.

Size of the corresponding intent, P_i, defines strength of the *p*-link.

Informally *p*-link between two organizations means that for a certain set of publishers P_i only these two institutions use them all, and vice versa, set of publishers used by the both institutions coincides with P_i.

See [16-18] for more details on using FCA technique for communities study.

2.4. Connectivity layers

In [3] it was suggested to classify all the nodes in *V* into the following connectivity layers:

- L_0 is the subset isolated nodes.
- L_1 denotes isolated nodes with external links.
- L_2 are members of pairs connected with collaborating institutions from the considered city.
- L_3 stands for dangling nodes belonging to a larger connected component.
- L_4 includes nodes on dangling paths.
- L_5 nodes from the graph 2-core.

To be more formal

$$L_5 = max\{U \subseteq V \mid \forall u \in U \exists v, w \in V (v \neq w \wedge a(u,v) \wedge a(v,w))\}.$$

Structural properties of segments $L_0 - L_4$ are rather trivial. Indeed, tree-like graph connectivity means that for any two vertices there exists a unique path linking them. From practical point of view such a structure may be considered as optimal from the prospective of information security. But in open academic community the absence of contacts usually implies not involvement into modern research, and as a result very modest scientific results. So, for the connectivity analysis of organizations in a city we calculate sizes of $L_0 - L_4$ and devote more profound study to L_5.

3. Data set

In this section the data that we used in our research is described.

3.1. Scientific Electronic Library, http://elibrary.ru

The publications data is available free of charge from the web site of Scientific Electronic Library, http://elibrary.ru. The license of the resource does not allow massive or systematic download of data, so we restricted our research to IT related articles which according to the standard Russian classification fall into two categories, Informatics and Cybernetics. Another restriction is time window: we analyzed articles published during 2011-2013. The focus of this study is restricted to Moscow and St. Petersburg, so the search filter was configured accordingly.

3.2. The data set characteristics

The search through Scientific Electronic Library returned information about 3992 papers for Moscow and 1485 papers for St. Petersburg. The graph size (number of organizations) for Moscow is 237, for St. Petersburg

is 110. Total number of links (including links to organizations in other cities, which are used to distinguish L_0 and L_1) is 501 for Moscow, and 184 for St. Petersburg.

4. Visualization of *a*-links

It is difficult to overestimate the role of good data visualization in the decision making process. Unfortunately, straightforward application of standard tools usually makes the resulting picture quite noisy and difficult to comprehend. It was already mentioned above that organization connectivity graphs for Moscow and St. Petersburg comprise hundreds of nodes and edges, so we tried different filters and pre-processing steps to figure out the most useful graph view. Finally, the following was done:

1. Displays a-links with strength 2 and more.
2. Do not display isolated nodes.
3. Use color and size indication to stress out nodes with higher centrality.

4.1. Moscow

The resulting graph of *a*-links for Moscow is given on Fig 1.

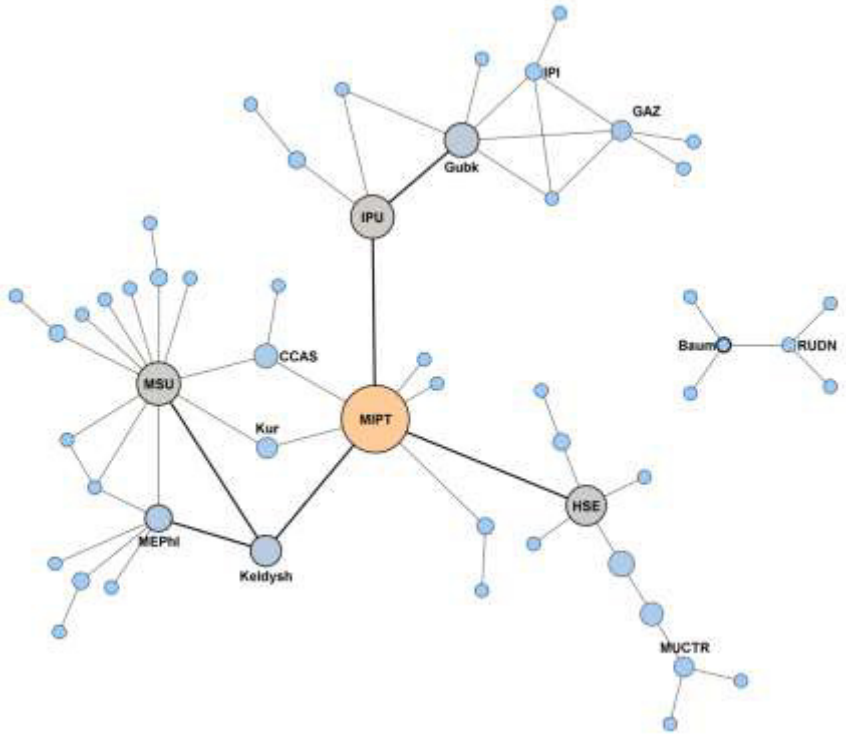

Fig. 1. Organizations connectivity graph for Moscow based on publications on Informatics and Cybernetics during 2011-2013 listed at Scientific Library portal, http://elibrary.ru

One can clearly see that Moscow Institute of Physics and Technology plays a central role in connecting different institutions in Moscow. Isolated island on the right is created by Bauman State Technical University and its partners.

4.2. St. Petersburg

The resulting graph of *a*-links for St. Petersburg is given on Fig 2.

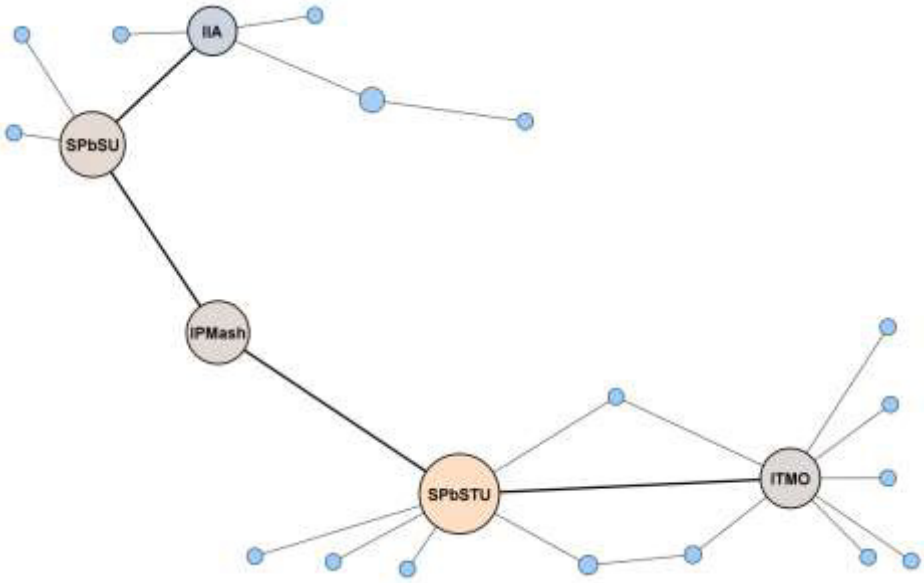

Fig. 2. Organizations connectivity graph for St. Petersburg based on publications on Informatics and Cybernetics during 2011-2013 listed at Scientific Library portal, http://elibrary.ru

4.3. Novosibirsk

Although in this paper we are focused at Moscow and St. Petersburg it is useful to compare their graphs to the similar one for Novosibirsk, which is given in Fig 3.

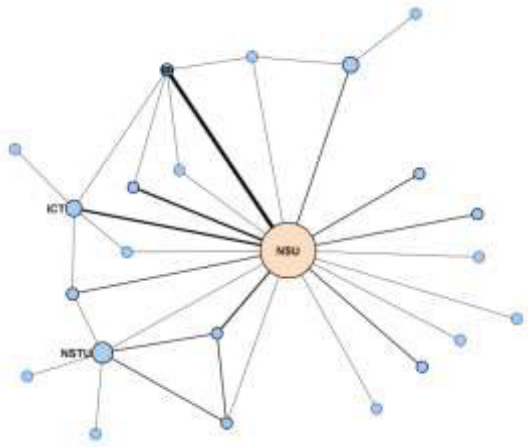

Fig. 3. Organizations connectivity graph for Novosibirsk from the similar dataset

5. Analysis of *p*-links

As visualization of a-links shows the current status of connectivity in a city, *p*-links could be used to figure out potential partners. First, we use the Concept Explorer tool to compute all formal concepts. Then, the following Python script was used to extract binary concepts from xml file (.cex) created by the tool, see Fig 4.

```
import xml.etree.ElementTree as ET
tree = ET.parse(inFileName)
root = tree.getroot()

lattices = root.find('Lattices')
for concept in lattices.iter('Intent'):
        if len(concept) == 2:
                for child in concept:
                        f2.write(child.get('AttributeIdentifier') + " ")
                f2.write("\n")
```

Fig. 4. Python script to extract binary concepts from xml file (.cex) created by Concept Explorer

The resulting graph of *p*-links for Moscow has 42 edges, that is organizations-publishers context has 42 binary concepts. Fragment of the lattice corresponding to the most publishing organizations according is given on Fig. 5.

Some of the results are quite surprising: As one can see from Fig. 1 there are no connections from HSE to MSU and MEPhI, which are definitely top international centers of computer science. Yet, in graph composed by p-links, HSE, MSU, MEPhI and Financial University form 4-clique, i.e. a sub-graph in which all pairs are connected.

From managerial point of view that could be interpreted as a subject for a specially designed program targeted at establishing better collaboration between these universities.

6. Conclusion

The research reported in this paper was originally initiated by Virtual Skolkovo program targeted at development of professional online community around Skolkovo Innovation Center, which is currently under construction in Moscow suburb. Recently it turned out that the developed methods and tools of measuring professional communities are in great demand by managers and supervisors of government and corporate programs of regional development.

In this paper an approach was suggested to measure connectivity between organizations in a given city. Openly available data on research publications is used. First kind of connections, *a*-links, help one to assess current level of collaboration in the city. The second, *p*-links, indicates potential partners.

The task of creating productive communities of practice in cities do not have a solution in today's understanding of sociology. Approach suggested in this paper is an attempt of interdisciplinary research, when the study is conducted on the basis of models of graph theory and sociological approach.

The study resulted in finding certain regularities in the life cycle of urban communities, which should be studied in detail in future.

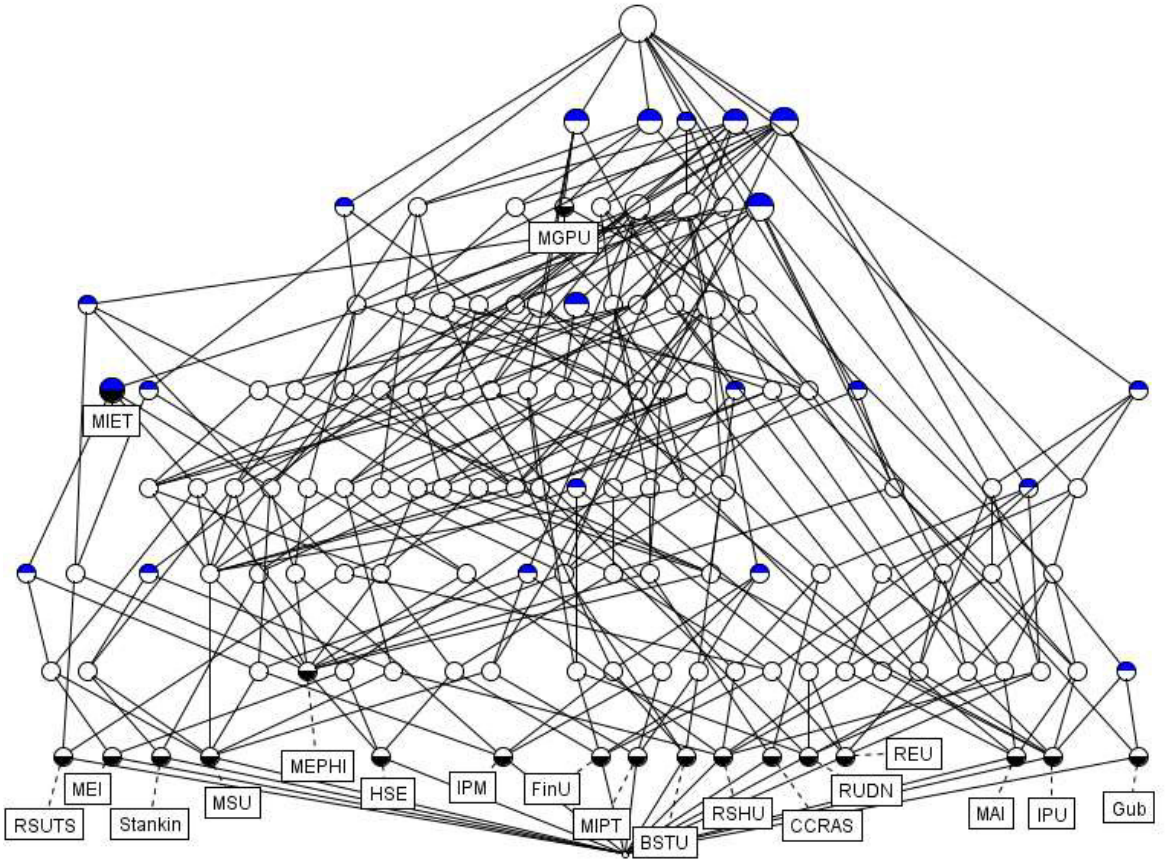

Fig. 5. The lattice of the Organizations-Publishers formal concepts

Acknowledgements

The authors express great appreciation to Oleg B. Alekseev, member of the Russian Ministry of Education and Science Council to improve the competitiveness of the leading universities of the Russian Federation (Government program 5/100) for his valuable and constructive suggestions during the planning and development of this research work.

References

[1] Fedor Krasnov, Rostislav Yavorskiy, “Measurement of maturity level of a professional community”, Business Informatics 2013. No. 1(23). pp 64–67 (in Russian)

[2] Fedor Krasnov, Dmitriy Ustalov, Rostislav Yavorskiy, “Comparison of Online Communities on the Base of Lexical Analysis of the News Feed”. Proceedings of 2-nd conference on Analysis of Images, Networks and Texts, Yekaterinburg, 4-6 april 2013, pp. 254-257. (in Russian)

[3] Fedor Krasnov, Evgeniya Vlasova, and Rostislav Yavorskiy, “Indicators of Connectivity for Urban Scientific Communities in Russian Cities”, Proceedings of 3-d international conference on Analysis of Images, Social Networks and Texts, Yekaterinburg, 10-12 april 2014.
[4] Andersson, David Emanuel, E. Andersson, and Charlotta Mellander, eds. *Handbook of creative cities*. Edward Elgar Publishing, 2011.
[5] Komninos, Nicos. *Intelligent cities: innovation, knowledge systems and digital spaces*. Routledge, 2013.
[6] Kienle, A., & Wessner, M. (2005). “Principles for cultivating scientific communities of practice. In Communities and Technologies” 2005 (pp. 283-299). Springer Netherlands.
[7] Kienle, A., & Wessner, M. (2006). “Analysing and cultivating scientific communities of practice”. International Journal of Web Based Communities, 2(4), 377-393.
[8] Tuire, Palonen, and Lehtinen Erno. "Exploring invisible scientific communities: Studying networking relations within an educational research community. A Finnish case." Higher Education 42.4 (2001): 493-513.
[9] Gaines, Brian R., and Mildred LG Shaw. "Using knowledge acquisition and representation tools to support scientific communities." AAAI. 1994.
[10] Smith, Marc A., and Peter Kollock, eds. Communities in cyberspace. Vol. 1. London: Routledge, 1999.
[11] Hildreth, Paul M., and Chris Kimble, eds. *Knowledge networks: Innovation through communities of practice*. IGI Global, 2004.
[12] Vázquez-Barquero, Antonio. *Endogenous development: Networking, innovation, institutions and cities*. Routledge, 2002.
[13] Gurstein, Michael, ed. *Community informatics: Enabling communities with information and communications technologies.* IGI Global, 2000.
[14] Ping, Guo, Chen Jianwei, and Song Zhihui. Construction of Evaluation System of Innovative City and Empirical Research." Journal of Applied Sciences 13.22 (2013).
[15] Scientific Electronic Library, http://eLibrary.ru
[16] B.Ganter, R.Wille. Formal Concept Analysis. Mathematical Foundations. Springer, 1999
[17] Sergei O. Kuznetsov, Sergei Obiedkov and Camille Roth, Reducing the Representation Complexity of Lattice-Based Taxonomies. In: U. Priss, S. Polovina, R. Hill, Eds., Proc. 15th International Conference on Conceptual Structures (ICCS 2007), Lecture Notes in Artificial Intelligence (Springer), Vol. 4604, pp. 241-254, 2007.
[18] Galitsky, Boris A., Boris Kovalerchuk, and Sergei O. Kuznetsov. "Learning common outcomes of communicative actions represented by labeled graphs." *Conceptual Structures: Knowledge Architectures for Smart Applications*. Springer Berlin Heidelberg, 2007. 387-400.