UDC 000.41+004.021

COMMUNITY STRUCTURE OF THE WORLD REVEALED BY FLICKR DATA

Alexander Belyi, Iva Bojic, Stanislav Sobolevsky, Lada Rudikova,
Alexander Kurbatski and Carlo Ratti

Recent availability of geo-localized data capturing individual human activity together with the statistical data on international migration opened up unprecedented opportunities for a study on global mobility. In this paper we consider it from the perspective of a complex network, built using a dataset of digital photos and videos posted on the Flickr website. This dataset provides insights on the global mobility highlighting short-term visits of people from one country to another. We use this mobility network to infer the structure of the global society through a community detection approach and demonstrate that consideration of mobility network between countries can reveal interesting global spatial patterns.

Introduction

People travel from one country to another for different reasons and while doing so, a lot of them leave their digital traces in various kinds of digital services. This opens tremendous research opportunities through the corresponding datasets, many of which have been already utilized for research purposes, including mobile phone call records [1-4], vehicle GPS traces [5-6], smart cards usage [7-8], social media posts [9-11] and bank card transactions [12-15].

Using these traces we can reconstruct people movements and afterwards analyze them to see if interesting or useful patterns emerge. Based on this results we can build models to predict where people will go next. It has already been shown that results of such analysis can be applied to a wide range of policy and decision-making challenges, such as for example regional delineation [16-17] or land use classification [18-19]. A number of studies focus specifically on studying human mobility at urban [20-22], country [23] or global scale. While considering aspects of human mobility at global scale, one can observe two major types of movements: an international migration [24-27] and short-term trips explored for example through geo-localized data from Twitter [28] or Flickr [29-30].

Some studies tried to primarily explain and model global mobility [24-27], while other rather focused on its applications, such as revealing the structure of the global society through global mobility networks [16, 17, 28]. Some scholars even considered relationships between human migration and economic links between countries [31-32]. However, global human dynamics by itself has a complex nature containing various types of mobility, including different processes as permanent relocation and short-term visits; thus it is extremely important to consider different aspects of human behavior. And analysis of the data from various available sources can help to achieve this.

In this study we use a dataset obtained from Flickr. Country-to-country mobility network extracted from it mostly represents short-term human mobility, as in the most cases it reflects activity during a leisure time and visits of touristic places. We explore the topological architecture of this global mobility network by studying its community structure, e. i., partitioning of the countries around the world into clusters. Detecting the community structure of the complex mobility network and understanding how it correlates with country-specific variables and geography is crucial from an international-travel perspective. Indeed, finding communities in the mobility network means identifying clusters of countries that carry tightly interrelated cultural and historical linkages among them, while being relatively less interconnected with countries outside the cluster. Following the approach from [16-18, 25] we use modularity maximization to determine the partitions. And to be able to deal with network without loop-edges we had to adjust modularity function. As final results we present and discuss partitions found at different levels of granularity obtained by performing clustering with different values of resolution parameter.

Dataset

A Flickr dataset used in our study contains more than 130 million photographs and videos. It was created by merging two publicly available Flickr datasets – one coming from a research project and another from Yahoo [33-34]. The records in two datasets partially overlap, but since each digital object in both datasets has its id, we were able to merge them by omitting duplicates and choosing only those records that were made within a 10 year time window, i.e., from 2005 and until 2014.

In order to build a directed and weighted network that describes short-term human mobility, we had to convert the Flickr datasets into origin-destination matrix where origins represented users' home countries and destinations are places (i.e., countries) where users took photos or created videos. Since the dataset does not contain information about user home location, we had to determinate it. Previous works showed that it is important to use proper method for home location definition [35], taking this into account we chose the most conservative method from techniques used in similar studies. Namely, we decided which of the users are acting in each location as residents by using the following criteria: a person is considered to be a resident of a certain country if this is the country where he/she took the highest number of the photographs/videos over the longest timespan (calculated as the time between the first and last photograph taken within the country) compared to all other countries for the considered person.

Using this simple criteria, we were able to determine home country for over 500 thousand users in the Flickr dataset that took almost 80% of all the photographs/videos in the dataset (i.e., more than 90 million in total), while the rest of the users for which home country could not be defined mostly belong to a low-activity group taking photographs only occasionally. When constructing weighted and directed mobility network, we only considered users for whom we were able to determine their home country. Finally, two countries are connected with a link if there is at least one person from the first country that had some activity in the second country where the value of every weighted link in this network corresponds to the total number of users from one country that made digital objects in the other one.

We should mention here that Flickr are much more widely used in developed countries while penetration into some other countries can be quite low. Figure 1 shows how many users per 1 million of population from each country we determined to be active outside their homeland. We can see that penetration in China and India as well as in most African counties is pretty low.

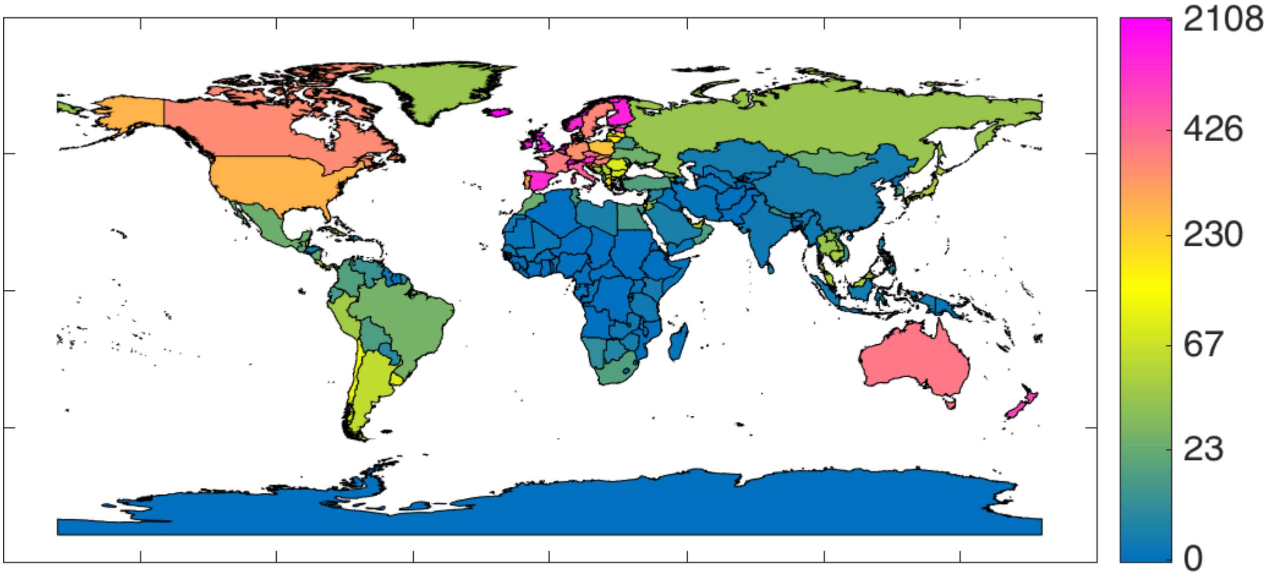

Fig. 1. Penetration of Flickr into countries all over the world as number of users who travel abroad per 1 million of population

Community detection method

Previous studies [16-17] have shown that community detection in human interaction and mobility networks usually leads to connected spatially cohesive communities (even with no spatial properties being considered in the community detection method) often revealing meaningful geographical patterns. There was no exception for the global mobility networks estimated from Twitter [28], as well as from migration data [25]. In this work we applied similar approach to the network based on the Flickr data.

We used one of the most popular and well established approaches to the network partitioning based on maximization of modularity function [36-37]. But in the mobility network we consider, loop edges are absent. In order to account for this we had to adjust classical modularity function. In particular, we altered the way null-model used by modularity estimates weight of each edge. In its classical form modularity uses $\frac{s_i t_j}{\sum_k s_k}$ as an expected weight of the edge from *i* to *j*, where if w_{ij} is the weight of the link from *i* to *j* then $s_i = \sum_j w_{ij}$ and $t_j = \sum_i w_{ij}$. This can be explained as that distribution of the outgoing weight s_i among all the possible destinations is proportional to their incoming weight t_j. However, if loop edges do not participate in this distribution then it should be rather $\frac{s_i t_j}{\sum_{k \neq i} s_k}$ or $\frac{s_i t_j}{\sum_{k \neq j} t_k}$, depending on whether it is seen as distribution of the outgoing weight s_i among all the destinations except *i* itself, or as distribution of the incoming weight t_j among all the origins except *j*. Finally, as a final estimation we use average of these, that leads to the expression $\frac{1}{2}\left(\frac{s_i t_j}{m - t_i} + \frac{s_i t_j}{m - s_j}\right)$, where $\mathrm{m} = \sum_k s_k = \sum_k t_k = \sum_{ij} w_{ij}$ is the total weight of all edges.

Since it has already been shown that modularity suffers from certain drawbacks, such as a resolution limit [38-39] preventing it from recognizing smaller communities, we also used the approach proposed by Arenas et al. [40] that involves introduction of a so-called resolution parameter, leading to the further adjustment of the modularity score. This way the final formula for the adjusted modularity measure used for our case of the mobility network free of the loop edges is:

$$\mathrm{Q}=\frac{1}{2m}\sum_{i\neq j}\left(2w_{\mathrm{ij}}-a\frac{s_i t_j}{m-t_i}-a\frac{s_i t_j}{m-s_j}\right)\delta\left(C_i,\mathrm{C}_j\right)$$

where *a* denotes the resolution parameter, *i*, *j* are nodes, c_i, c_j – the communities they belong to, $\delta(\mathrm{x,y}) = 1$ if x=y, *0* otherwise.

Finally, in order to find the best partitioning, we optimized this version of modularity using efficient and precise Combo algorithm [41], suitable for dealing with different types of objective functions.

Community structure of the mobility network

For the sake of noise reduction, we excluded nodes for which incoming or outgoing strength was less than 10, that left us with a network of 201 countries. We consider partitioning for different values of resolution parameter. Applying modularity maximization with the default resolution parameter of 1.0 leaves us with only *five* communities, while for resolution parameter equal to 2.0, number of obtained communities goes up to *seventeen* making it already harder to visually recognize and analyze different communities on a map. That is why we considered only partitions for parameter taking values between 1.0 and 2.0 presenting the results in Figures 2, 3 and 4 for resolution parameter values 1.0, 1.5 and 2.0, respectively (countries are colored according to the community they belong to).

From all figures we can see that main geographical regions such as North and South Americas, East Asia, South Africa, Commonwealth of Independent States (CIS) are usually united into one community. Partitions start to be more complicated for higher *a*, but this is because more local patters are discovered. Nevertheless, we can see some interesting features at each level. For example, Egypt and Turkey are always fall into the same community with most CIS countries which could be explained by (and can serve as a nice evidence of) high popularity of these counties as destinations for tourists from CIS. Looking at community structure of higher granularity can reveal significant presence of US in Iraq and Afghanistan as well as still strong relationships between European countries and their African ex-colonies. At the same time only Ireland falls into community of United Kingdom, once mighty dominion with colonies spread across the whole globe. Found mobility clusters intimate that while people tend to travel more to close-by destinations rather than further afield, common language and history play important role in choice of travel destination.

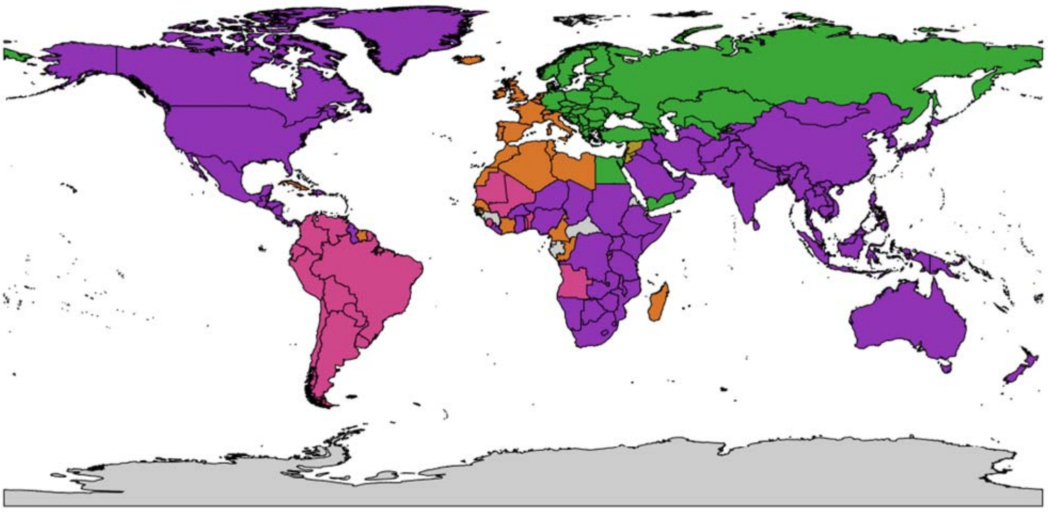

Fig. 2. Community structure for resolution parameter value equal to 1.0

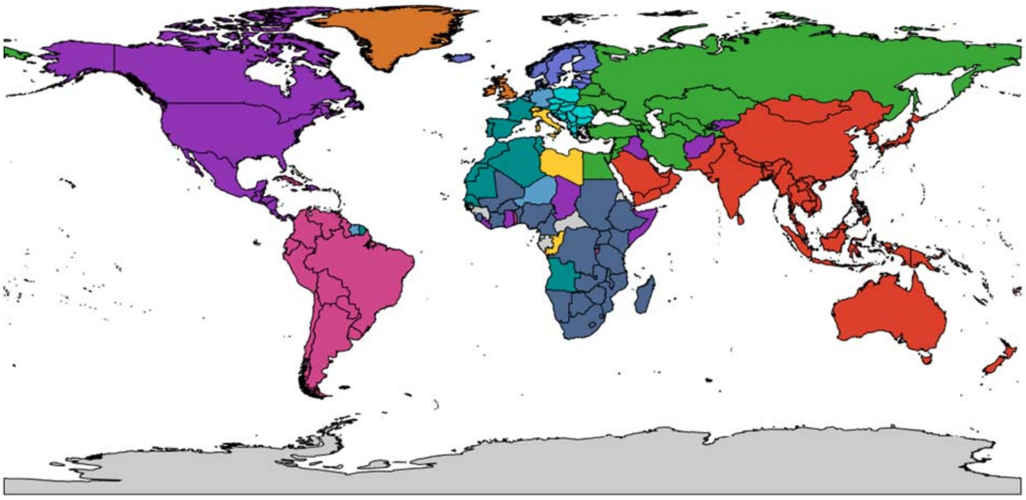

Fig. 3. Community structure for resolution parameter value equal to 1.5

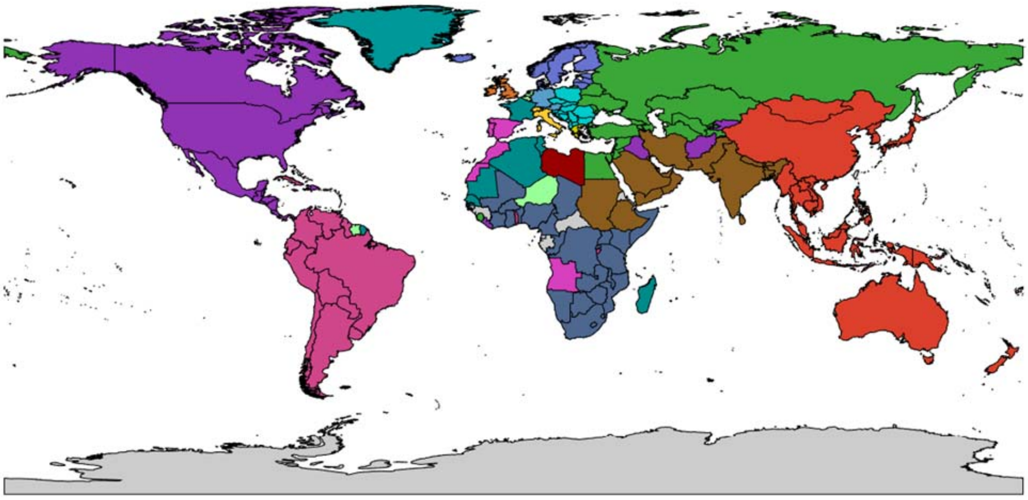

Fig. 4. Community structure for resolution parameter value equal to 2.0

From the results shown it is clear that found community structure for mobility network constructed on Flickr data has interesting property inherent in most other mobility networks studied: clusters are spatially continuous and reflect common world regions. This is consistent with the previous findings for mobility networks constructed using mobile phone [16-18], Twitter [28] and migration [25] data. The reasons behind grouping some countries into one community can include close geographical distance, strong economic ties and cultural aspects.

Conclusions

In this study we explored short-term human mobility network created using dataset of media objects posted on the Flickr website. In order to study the structure of the global human society we applied a community detection method to this network. While applying well established modularity maximization approach, we improved it by altering underlying null-model to better account for the absence of loop-edges in network. After consideration of different values of resolution parameter, we obtained partitions with higher granularity. We discussed results pointing specific spatial patterns revealed. Also by this work, we confirmed finding of previous studies that partitioning of mobility network results in meaningful geographically connected communities.

References

1. Ratti, C. Mobile landscapes: Using location data from cell phones for urban analysis / C. Ratti [et al.] // Environment and Planning B. – 2006. – Vol. 33. – P. 727–748.
2. Calabrese, F. Real time rome / F. Calabres, C. Ratti // Networks and Communication Studies. – 2006. – Vol. 20. – P. 247–258.
3. Girardin, F. Digital footprinting: Uncovering tourists with user-generated content / F. Girardini [et al.] // IEEE Pervasive Computing. – 2008. – Vol. 7 – P. 36–43.
4. Quercia, D. Recommending social events from mobile phone location data / D. Quercia [et al.] // in Proceedings of the 10th IEEE International Conference on Data Mining. – IEEE, 2010. – P. 971–976.
5. Santi, P. Quantifying the benefits of vehicle pooling with shareability networks / P. Santi [et al.] // Proceedings of the National Academy of Sciences. – 2014. – Vol. 111. – P. 13290–13294.
6. Kang, C. Exploring human movements in Singapore: A comparative analysis based on mobile phone and taxicab usages / C. Kang [et al.] // in Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing. – ACM New York. – 2013. – P. 1–8.
7. Bagchi, M. The potential of public transport smart card data / M. Bagchi, P. White // Transport Policy. – 2005. – Vol. 12. – P. 464–474.
8. Lathia, N. The hidden image of the city: Sensing community well-being from urban mobility / N. Lathia, D. Quercia, J. Crowcroft // Pervasive Computing. – Springer, 2012. – Vol. 7319 of Lecture Notes in Computer Science. – P. 91–98.

9. Java, A. Why we twitter: understanding microblogging usage and communities / A. Java [et al.] // in Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. – ACM, New York, 2007. – P. 56–65.
10. Szell, M. Contraction of online response to major events / M. Szell, S. Grauwin, C. Ratti // PloS One. – 2014. – Vol. 9. – P. 1–9.
11. Frank, M. R. Happiness and the patterns of life: A study of geolocated Tweets / M. R. Frank [et al.] // Scientific Reports. – 2013. – P. 1–9.
12. Sobolevsky, S. Mining urban performance: Scale-independent classification of cities based on individual economic transactions / S. Sobolevsky [et al.] // in Proceedings of the 2nd ASE International Conference on Big data Science and Computing. – ASE, Stanford, 2014. – P. 1–10.
13. Sobolevsky, S. Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. The case of residents and foreign visitors in Spain / S. Sobolevsky [et al.] // in Proceedings of the IEEE International Congress on Big Data. – IEEE, 2014. – P. 136–143.
14. Sobolevsky, S. Cities through the prism of people's spending behavior / S. Sobolevsky [et al.] // arXiv preprint arXiv:1505.03854. – 2015. –P. 1–21.
15. Sobolevsky, S. Predicting regional economic indices using big data of individual bank card transactions / S. Sobolevsky [et al.] // in Proceedings of the 6th ASE International Conference on Data Science. – ASE, Stanford, 2015. – P. 1–12.
16. Ratti, C. Redrawing the map of Great Britain from a network of human interactions / C. Ratti [et al.] // PLoS One. – 2010. – Vol. 5. – P. 1–6.
17. Sobolevsky, S. Delineating geographical regions with networks of human interactions in an extensive set of countries / S. Sobolevsky [et al.] // PloS One. – 2013. – Vol. 8. – P. 1–10.
18. Pei, T. A new insight into land use classification based on aggregated mobile phone data / T. Pei [et al.] // International Journal of Geographical Information Science. – 2014. – Vol. 28, 1988–2007.
19. Grauwin, S. Towards a comparative science of cities: Using mobile traffic records in New York, London and Hong Kong / S. Grauwin [et al.] // in Computational Approaches for Urban Environments. – Springer, 2015. – Vol. 13 of Geotechnologies and the Environment. – P. 363–387.
20. Gonzalez, M. Understanding individual human mobility patterns / M. Gonzalez, C. Hidalgo, A.-L. Barabasi // Nature. – 2008. – Vol. 453, – P. 779–782.
21. Kung, K. Exploring universal patterns in human home/work commuting from mobile phone data / K. Kung [et al.] // PLoS One. – 2014. – Vol. 9. – P. 1–15.
22. Hoteit, S. Estimating human trajectories and hotspots through mobile phone data / S. Hoteit [et al.] // Computer Networks. – 2014. – Vol. 64. – P. 296–307.
23. Amini, A. The impact of social segregation on human mobility in developing and industrialized regions / A. Amini [et al.] // EPJ Data Science. – 2014. – Vol. 3. – P. 1–20.

24. Greenwood, M. J. Human migration: Theory, models, and empirical studies* / M. J. Greenwood // Journal of regional Science. – 1985. – Vol. 25. – P. 521–544.
25. Fagiolo, G. International migration network: Topology and modeling / G. Fagiolo, M. Mastrorillo // Physical Review E. – 2013. – Vol. 88. – P. 012812.
26. Abel, G. J. Quantifying global international migration flows / G. J. Abel, N. Sander / Science. – 2014. – Vol. 343. – P. 1520–1522.
27. Tranos, E. International migration: a global complex network / E. Tranos, M. Gheasi, P. Nijkamp // Environment and Planning B: Planning and Design. – 2015. – Vol. 42. – P. 4–22.
28. Hawelka, B. Geo-located Twitter as proxy for global mobility pattern / B. Hawelka [et al.] // Cartography and Geographic Information Science. – 2014. – Vol. 41. – P. 260-271.
29. Paldino, S. Urban magnetism through the lens of geo-tagged photography / S. Paldino [et al.] //EPJ Data Science. – 2015. – Vol. 4, – P. 1–17.
30. Sobolevsky, S. Scaling of City Attractiveness for Foreign Visitors through Big Data of Human Economical and Social Media Activity / S. Sobolevsky [et al.] // in Big Data (BigData Congress), 2015 IEEE International Congress on – IEEE, 2015. – P. 600–607.
31. P. Sgrignoli, R. Metulini, S. Schiavo, and M. Riccaboni, Physica A: Statistical Mechanics and its Applications 417, 245 (2015).
32. Fagiolo, G. Does human migration affect international trade? A complex-network perspective / G. Fagiolo, M. Mastrorillo // PLoS ONE. – 2014. – Vol. 9. – P. e97331.
33. Sfgeo.org [Electronic resource] / – San Francisco, 2010. – Mode of access: http://sfgeo.org/data/tourist-local. – Date of access: 1.04.2016.
34. Thomee, B. The New Data and New Challenges in Multimedia Research / B. Thomee [et al.] // arXiv preprint arXiv:1503.01817 – 2015.
35. Bojic, I. Choosing the Right Home Location Definition Method for the Given Dataset / I. Bojic [et al.] // Social Informatics. – Springer International Publishing, 2015. – Vol. 9471 in Lecture Notes in Computer Science. – P. 194–208.
36. Newman, M. Finding and evaluating community structure in networks / M. Newman, M. Girvan // Physical Review E. – 2004. – Vol. 69. – P. 026113.
37. Newman, M. Modularity and community structure in networks / M. Newman // Proceedings of the National Academy of Sciences. – 2006. – Vol. 103. – P. 8577–8582.
38. Fortunato, S. Resolution limit in community detection / S. Fortunato, M. Barthelemy // Proceedings of the National Academy of Sciences. – 2007. – Vol. 104. – P. 36–41.
39. Good, B. H. Performance of modularity maximization in practical contexts / B. H. Good, Y.-A. de Montjoye, A. Clauset // Physical Review E. – 2010. – Vol. 81. – P. 046106.

40. Arenas, A. Analysis of the structure of complex networks at different resolution levels / A. Arenas, V. Fernandez, S. Gomez // New Journal of Physics. – 2008. – Vol. 10. – P. 053039.
41. Sobolevsky, S. General optimization technique for high-quality community detection in complex networks / S. Sobolevsky [et al.] // Physical Review E. – 2014. – Vol. 90. – P. 012811.

Alexander Belyi, PhD student at Belarusian State University faculty of applied mathematics and computer science and software engineer at Singapore MIT Alliance for Research and Technology Centre, alex.bely@smart.mit.edu

Iva Bojic, Postdoctoral associate at Singapore MIT Alliance for Research and Technology Centre, PhD in Computer Science, ivabojic@smart.mit.edu

Stanislav Sobolevsky, Associate professor of practice at the Center for Urban Science and Progress at New York University and a research affiliate at the MIT Senseable City Lab, Doctor of Physical and Mathematical Sciences, Professor, sobolevsky@nyu.edu

Lada Rudikova, Chair of department of intelligent software and computer systems at Yanka Kupala State University of Grodno, PhD, Associate professor, rudikowa@gmail.com

Alexander Kurbatski, Chair of department of software engineering at Belarusian State University, Doctor of Technical Sciences, Professor, kurb@unibel.by

Carlo Ratti, Director of MIT Senseable City Lab, PhD, professor, ratti@mit.edu