AASCIT Communications
Volume 3, Issue 3
March 4, 2016 online
ISSN: 2375-3803

American Association for Science and Technology

Big Data Processing in Complex Hierarchical Network Systems I: Structures and Information Flows

Olexandr Polishchuk Department of Nonlinear Mathematical Analysis, Pidstryhach Institute for Applied Problems of Mechanics and Mathematics, National Academy of Sciences of Ukraine, Lviv, Ukraine

Dmytro Polishchuk Department of International Communications, Computer Center of Lviv Railway, Lviv, Ukraine

Maria Tyutyunnyk Department of Nonlinear Mathematical Analysis, Pidstryhach Institute for Applied Problems of Mechanics and Mathematics, National Academy of Sciences of Ukraine, Lviv, Ukraine

Mykhailo Yadzhak Department of Nonlinear Mathematical Analysis, Pidstryhach Institute for Applied Problems of Mechanics and Mathematics, National Academy of Sciences of Ukraine, Lviv, Ukraine

Keywords

Complex System, Network, Multiplex, Association, Conglomerate, Continuous Monitoring, Big Data, Data Processing

This article covers the problem of processing of Big Data that describe process of complex networks and network systems operation. It also introduces the notion of hierarchical network systems combination into associations and conglomerates alongside with complex networks combination into multiplexes. The main types of information flows in complex hierarchical network systems being the basic components of associations and conglomerates are covered.

Introduction

Complex large scale technological systems (CLSTS) are used almost in all areas of human activity, e.g. in transportation (railway, road and aviation systems, transportation networks of large cities and regions of countries) [1], supply and logistics (systems for power, gas, petrol, heat and water supply, trade networks) [2], information and communication (Internet, TV, radio, post service, press, fixed and mobile telephony) [3], in economics (networks of state-owned and (or) private companies, their suppliers and final products distributors) [4], finance (banking and insurance networks, money transfer systems) [5], education, healthcare etc. Their state and operation quality impose large impact on citizens' quality of life, efficiency of economy and possibilities for its development, as well as government structures readiness to mitigate impacts of technological and natural disasters. Finally, they may be treated as the evidences of country development level in general [6, 7]. Failure of one of the elements of such system can often lead to operation breakdown or destabilization of the whole CLSTS. The example of this is cascading phenomenon [8]. Often the situations of the kind (e.g. accidents at nuclear or large chemical plants and other hazardous facilities, power lines, gas pipelines etc.) may lead to harsh consequences, such as environmental disasters, property loss and numerous human victims [9]. These circumstances determine the importance of continuous monitoring of technological systems operation, careful control of their behavior and timely response to emerging threats [10]. In this case, the major problem consists in the need to analyze large amounts of data that describe the state and behavior of CLSTS elements. This problem can be solved through creation of efficient computing environments, and usage of applied systems analysis and artificial intelligence methods [11, 12]. These methods implement effective algorithms for processing and analysis of information arriving from system components.

Complex Networks, Network Systems, Multiplexes, Associations and Conglomerates

During recent years, the theory of complex networks has been rapidly developing [13, 14]. We encounter network structures [15, 16] while studying micro- (e.g. quantum networks of fermions connected with bosons) and macroworld (gas networks in Universe, networks of black holes etc.). They occur in nature (e.g. protein and metabolism networks) and human society (e.g. Internet, language and citation networks). Complex technology systems (transportation, trade and power supply networks etc.) are not exceptions either. In general, an arbitrary network is defined as a statistical assembly, i.e. a set of networks with each network having certain probability of implementation, or as a set of all possible conditions of the given network. On the other hand, complex networks are graphs, i.e. the sets of nodes connected by some relations with nontrivial topological properties. When talking about real networks these properties determine network operation features [17].

Nodes of one network may be the nodes of many other networks at the same time. Thus, every town in the country can be a node for several transportation networks, as well as state and local administration networks, economy and financial network etc. Every person is also the node of many networks (family, professional, social etc.). Combination of several networks with non-empty intersection of nodes is called a multiplex [18]. Each network being the component of multiplex is called a layer. Examples above show that there are different types of interactions between the nodes existing on different layers of multiplex. These interactions may be of various nature or meaning and may have different material media.

Sometimes complex networks are called "the systems". In our opinion, network only reflects the structure of a system being its frame. There are flows that move along the network that make it a system. Real networks are created and exist with an aim to arrange the flows of certain type. These flows can be continuous (e.g. power resources), discrete (e.g. trains) or continuous-discrete (phone calls). The motion of flows in the network can be ordered (railway traffic), partially ordered (car traffic in large cities) or unordered (information flows in social networks). Networks with different types and levels of flow arrangement are generated different network systems [19]. Flow properties of certain network allow to divide multiplexes into network layers in a proper way.

CLSTS complies with any definition of a complex network and even network system only to some extent. The reason is that flows movement in majority of artificial technology networks need to be supported on organizational level. This function is performed by CLSTS control system which has hierarchical structure. Hierarchical network structures (HNS) are special because each subsystem of a certain hierarchy level consists of a set of subsystems which form subnetwork of lower hierarchy level network (see Fig. 1a). Flow movement for which the network was created is performed at network of the lowest level. At the higher (control) levels, flows are represented by information, organizational and administrative decisions etc. HNS differs from common three dimensional tree structure by links between the nodes of each hierarchy level.

There is an alternative to multiplex method of complex network systems combination. It consists in their joint engagement for solving important social or economical problems, e.g. industrial or natural disasters, pandemics, acts of terrorism etc. Solving these problems requires the interaction of many systems of different type and purpose: rescue and fire services, police, security agencies, military and medical units etc. The structures of the kind often arise in the industry, politics, social life etc. We call supersystems resulting from diverse complex hierarchical network systems (CHNS) interaction “conglomerates”. Conglomerates often play more important role than multiplexes. Effective operation of conglomerates requires timely information exchange between their components. First of all, this means that interaction between CHNSs composing conglomerate has to be very tight. Relations between conglomerate components get less or more tighter depending on the purpose of their interaction and the extent of its implementation. Existence of purpose of creation and specific nature of relations are another aspects that make this structure different from multiplex.

The components of the conglomerate can comprise combination of several interacting systems of the same type. We call the structures of the kind "associations". Signs of uniformity determine whether some formation is a conglomerate or an association. If we talk about economy of the country in general, the combination of all transportation networks can be considered an association, since in this case the sign of uniformity is represented by the purpose of such system operation which lies in passengers and cargo transportation. At the same time, transportation system of the country is conglomerate of transportation systems of different types. Associations among components of such conglomerates are presented by transport operators organizations, transportation companies etc.

Sometimes the dilemma arises of whether some formation shall be considered a multiplex or a conglomerate. In general, it depends on purpose and extent of the study. Network of towns within the country is the basis for organization of flows of various types. From this point of view, it can be considered multiplex each layer of which provides a different type of flow movement (rail, road, air, sea, river), and each node of multiplex can support movement of up to all possible types of flows.

However, the cooperation of transport systems of various types with an aim to transport passengers and cargo inside and outside the country can be considered conglomerate. Thus, the transportation system of the country can be considered multiplex, conglomerate or association within the larger conglomerate (for example, industrial) depending on the purpose of the study. Note that during the study of complex networks and multiplexes, structure properties have higher priority, and during the study of network systems, associations and conglomerates, the most important is the function implementing the purpose of their creation.

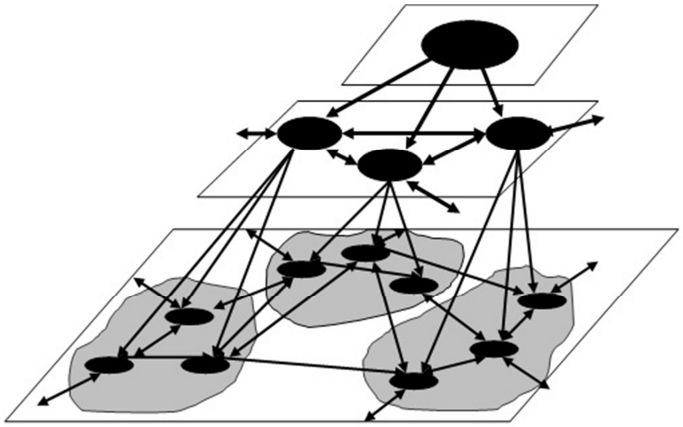


Fig. 1a. *Hiererchical network structure.*

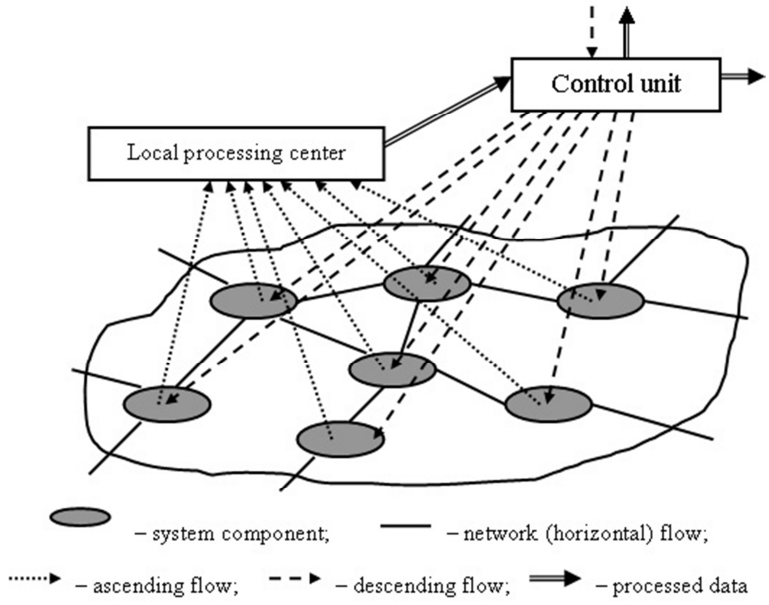


Fig. 1b. *Subsystem of CHNS.*

In general, the modern world is a huge dynamic multiplex-conglomerate structure with thousands of systems and billions of elements (nodes). Complex study of real multiplexes and conglomerates that constitute components of this structure requires complete and comprehensive understanding of their state and operation even in the case of their decomposition into network layers, associations and separate network systems. Such understanding can be achieved with the use of information about the history, current condition and forecast of systems behavior. The volume of this information, its diversity and problem of timely information flows content analysis lead to emergence of the phenomenon called Big Data.

Big Data and Information Overload

Big Data is the series of approaches and tools for processing huge volumes of structured and unstructured data. Properties of Big Data is defined by “three V”: volume (large volume of data), variety (diversity of data), and velocity (processing speed and efficiency of results obtaining) [20]. The main goal of Big Data processing is to achieve the results that can be perceived by

human. Before making decision, people try to get information that provides comprehensive characteristics of the problem being solved. However, as well as the lack of information, the excess of it can lead to achieving wrong results. Such situation is called information overload or the problem of analysis paralysis. It applies to CLSTS as well. In 1995, the study was carried out in the USA [21] with an aim to calculate the number of documents being processed by the employees of three large corporations. The result achieved was between 35 and 64 thousands. 20 years passed since that time. However, it is unlikely that information load on people working in those corporations, as well as other ones, has decreased since the volume and the number of information flows rise every year. It is important that the person responsible for operation of critical and dangerous facilities does not get lost in the flow of data describing the operation process of those facilities. Their task is to timely localize and solve the problem. To overcome information overload it is necessary to extract only the information essential for making correct and timely decision. Indeed, only small amount of data available to the person making decision satisfies this requirement. For example, the cracks in the railway rail have often led to accidents with numerous victims. The railway detector carriage used for cracks detection extracts the data with 1 mm step (in many countries, railroads are tens and hundreds of thousands kilometers long). As the result, multibillion arrays of numerical data [22] are obtained. At the same time, the only thing end user (maintenance unit) must know is the exact location of the crack. Many examples alike may be provided from power industry, economy and finance, medicine etc. The data regarding system elements behavior may continuously come in huge amounts from many sources and require real time processing. It is often necessary to store these data for long and short term CLSTS state and behavior forecasting. Many systems are sensitive to small changes the accumulation of which can be a threat to their normal operation. Analysis of stored information allows to identify the negative phenomena trends in advance and to prevent them beforehand.

The greatest problem of Big Data processing is induced by so-called unstructured information (text, photo, audio and video files). However, most of these data can be easily recognized and sorted by type, format, creation method etc. and sent to the relevant specialized tools for processing and analysis. Henceforth we will focus on processing of structured (in particular, numerical) data that usually make up the largest part of objective information about CLSTS operation process.

Information Flows in Complex Hierarchical Network Systems

We consider conglomerates as the combination of CLSTSs that interact to achieve common goals. Some CLSTSs can be grouped into associations according to uniformity. Every CLSTS is a complex hierarchical network system. At each level of the hierarchy, the edges ensure smooth motion of flows of certain type, whereas the nodes ensure their processing. Under the component of the system we understand its structural unit of any hierarchy level from element to subsystem of the highest decomposition level. Subsystems of the lowest decomposition level that consist of elements will be regarded to as the “basic subsystems” (BSS). The principles of the complex hierarchical network systems operation and methods of their behavior analysis are described in detail in [1].

In CHNS three main types of flows are distinguished (see Fig. 1b):

1) ascending flows which come from controlled components to control components and may contain either processed primary data or aggregated data;

2) descending flows which come from control components to controlled components and contain information necessary for normal operation of controlled components, as well as decisions regarding their further actions;

3) network (horizontal) flows that come from some network components of certain hierarchy level to other components of the same hierarchy level.

In general, ascending and descending flows implement cyclic (reverse) connection between control and controlled system components. These flows facilitate making correct and timely decisions regarding CLSTS operation process. Network flows provide self-organization of this process at every hierarchy level. These are the messages from one station regarding train delay to other stations located on the route of this train. Another example of such flows is professional information exchange in a team working on some project. In addition to the above mentioned, there are flows which run between the interacting systems of association or conglomerate. That is, we can add the intersystem flows to the listed intrasystem flows. One of the most important problems of CLSTS information support is synchronization of data flows between different CHNS hierarchy levels, within the networks of each hierarchy level and between the components of association or conglomerate.

We can also distinguish different levels of information processing. The first (the lowest) level consists in continuous extracting and real time processing of large volumes of numerical data that simultaneously arrive from many network elements (nodes) to local data processing and control centers. At this hierarchy level (level of system BSS) data analysis and timely

response to detected local problems in the network elements (nodes, edges and flows) is carried out. The main purpose of information analysis at this level is to discover potentially failure (or "the weakest") BSS elements. Failure of such elements often leads to the cascading phenomena (all catastrophes usually start from "minor failures" at system elements level). In such cases response time is major factor that allows to timely localize the problem and quickly overcome its consequences. Here it is expedient not only to determine current state but also to predict occurrence of potentially failure elements on the basis of previous evaluations history. In case such elements are missing, information processing center prepares the aggregated reports regarding condition of BSS elements and BSS in general for relevant control units. These reports are used as the basis for making objective and reasoned decision regarding further actions on this BSS. On this level, data on non-critical negative (which over time tends to become critical one) are accumulated for submission to upper control levels with an aim to solve existing problems and to prevent potential ones. In fact, at this level, objective information used to support decision-making by the control system units of the highest hierarchy levels shall be extracted and ordered.

Higher hierarchy levels receive the data aggregated in different ways [23]. Basic requirements for this data are as follows:

1) objectivity, i.e. they shall be based on reliable information only;

2) comprehensive description of the situation;

3) minimum sufficiency, i.e. the absence of duplicate and unimportant data;

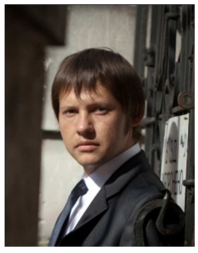
4) simplicity and understandability, i.e. visualization methods shall be chosen that allow to quickly orient in the huge amount of conclusions obtained.

Conclusions

Complex large scale technological systems of various types and destination are important in the life of society. In this paper we have considered the basic types of CLSTS structures, goals and ways of combining them into a more global structures: multiplexes, associations and conglomerates. The functioning of complex systems is accompanied by generation of the huge volumes of information, which at the time of decision-making can lead to the problem of analysis paralysis. We have identified the main types of data flows in hierarchical network structures, formulated requirements for this information and problems of their processing. For solving of these problems in the second part of this paper we will propose approaches for creation of efficient computing environments and distributed computations organization. Methods for preprocessing of large data sets and procedures for parallelization of primary data processing and formation of generalized conclusions at different levels of system hierarchy also will be described. ■

Olexandr Polishchuk

Olexandr Polishchuk is senior researcher at the Department of Nonlinear Mathematical Analysis of Pidstryhach Institute of Applied Problems of Mechanics and Mathematics, National Academy of Sciences of Ukraine, Lviv, Ukraine. He holds a M.Sc. in Applied Mathematics and a Ph.D. in Calculus Mathematics from the Computer Centre of Siberian Division of Academy of Sciences of USSR. His research interests include Complex Networks and Network Systems, Data Analysis and Data Mining, Artificial Intelligence, Evaluation Theory and Optimal Control. His research work has been published in many international and national journals and from conference proceedings. Now he is a member of AASCIT.
od_polishchuk@ukr.net



Dmytro Polishchuk

Dmytro Polishchuk is leading engineer at the Department of International Communication of Computer Center of Lviv Railway, Lviv, Ukraine. He holds a M.Sc. in Mathematical Statistics from the Lviv National University. His research interests include Complex Networks and Network Systems, Data Analysis and Data Mining, Statistic and Evaluation Theories. His research work has been published in many international and national journals and from conference proceedings.
d_pole@mail.ru

Maria Tyutyunnyk

Maria Tyutyunnyk is researcher at the Department of Nonlinear Mathematical Analysis of Pidstryhach Institute of Applied Problems of Mechanics and Mathematics, National Academy of Sciences of Ukraine, Lviv, Ukraine. He holds a M.Sc. in Applied Mathematics from the Lviv National University. Her research interests include High-Performance Computing Environments and Parallelization, Data Analysis and Evaluation Theory. Her research work has been published in many international and national journals and from conference proceedings.
dept25@iapmm.lviv.ua

Mykhailo Yadzhak

Mykhailo Yadzhak is leading researcher at the Department of Nonlinear Mathematical Analysis of Pidstryhach Institute of Applied Problems of Mechanics and Mathematics, National Academy of Sciences of Ukraine, Lviv, Ukraine. He holds a M.Sc. in Pure Mathematics, Ph.D. and D.Sc in Mathematical Programming from the Kyiv National University. His research interests include High-Performance Computing Environments and Parallelization, Data Analysis and Evaluation Theory. His research work has been published in many international and national journals and from conference proceedings.
yadzhak_ms@ukr.net

References

[1] Polishchuk D., Polishchuk O., and Yadzhak M. “Complex evaluation of hierarchically-network systems”, *Automatic Control and Information Sciences,* vol. 2, no. 2, pp. 32-44, 2014.

[2] Neng Chiu H., “The integrated logistics management system: a framework and case study”, *International Journal of Physical Distribution & Logistics Management*, vol. 25, no. 6, pp. 4-22, 1995.

[3] Berners-Lee T. et all, “World-Wide Web: the information universe”, *Internet Research*, vol. 20, no. 4, pp. 461-471, 2010.

[4] Marsan G. A., Bellomo N., and Egidi M., “Towards a mathematical theory of complex socio-economical systems by functional subsystems representation”, *Kinetic and Related Models*, vol. 1, no. 2, pp. 249-278, 2008.

[5] Rajan R. G. and Zingales L., “Financial systems, industrial structure, and growth”, *Oxford review of economic Policy*, vol. 17, no. 4, pp. 467-482, 2001.

[6] Scott A. J. and Storper M. “Regions, globalization, development”, *Regional studies*, vol. 41 (S1), pp. 191-205, 2007.

[7] Polishchuk O., Polishchuk D., Tyutyunnyk M., Yadzhak M. “Issues of regional development and evaluation problems”, *AASCIT Communications*, vol. 2, no. 3, pp. 115-120, 2015.

[8] Buldyrev S. et all. “Catastrophic cascade of failures in interdependent networks”, *Nature,* vol. 464(15), pp. 1025–1028, 2010.

[9] Weart S. R. and Weart S. R. *Nuclear fear: A history of images*, Harvard University Press, 2009.

[10] Blanchard B. S. and Fabrycky W. J., *Systems engineering and analysis (Vol. 4),* Englewood Cliffs, New Jersey: Prentice Hall, 1990.

[11] Isermann R., Fault-diagnosis applications: model-based condition monitoring: actuators, drives, machinery, plants, sensors, and fault-tolerant systems, Springer Science & Business Media, 2011.

[12] Tong C. and Sriram D., Artificial Intelligence in Engineering Design: Volume III: Knowledge Acquisition, Commercial Systems, And Integrated Environments, Elsevier, 2012.

[13] Barabási A.-L., Frangos J. *Linked: the new science of networks science of networks*, Basic Books, 2014.

[14] Boccatti S. et all. “Complex networks: structure and dynamics”, *Physics Reports*, vol. 424, pp. 175-308, 2006.

[15] Bornholdt S. and Schuster H. G. *Handbook of Graphs and Networks: From the Genome to the Internet,* Jon Wiley & Sons, 2006.

[16] Dorogovtsev S. N. and Mendes J. F. F. *Evolution of Networks: From Biological Nets to the Internet and WWW*, Oxford University Press, 2013.

[17] Newman M., Barabasi A.-L., and Watts D. J. *The structure and dynamics of networks*, Princeton University Press, 2006.

[18] Babak F. and Naghmeh M. “Growing Multiplex Networks with Arbitrary Number of Layers”, arXiv:1506.06278v1 [physics.soc-ph], 20 Jun 2015.

[19] Polishchuk D., Polishchuk O., and Yadzhak M. “Complex deterministic evaluation of hierarchically-network systems: I. Methods description”, *System Research and Information Technologies*, vol. 1, pp. 21-31, 2015.

[20] Mayer-Schönberger V. and Kenneth C. *Big data: A revolution that will transform how we live, work, and think*, Houghton Mifflin Harcourt, 2013.

[21] UNESCO Statistical Year-book, 1995.

[22] Polishchuk, D. O. "Evaluation of ukrainian railway equipment condition", *Science and Transport Progress. Bulletin of Dnipropetrovsk National University of Railway Transport*, vol. 41, pp. 203-211, 2012.

[23] Polishchuk D., Polishchuk O., and Yadzhak M. “Complex deterministic evaluation of hierarchically-network systems: I. Aggregated evaluation”, *System Research and Information Technologies*, vol. 4, pp. 20-31, 2015.