# Modelling and Analysis of Complex Patient-Treatment Process using GraphMiner Toolbox

Oleg Metsker[1], Sergey Kesarev[1], Ekaterina Bolgova[1], Kirill Golubev[1],
Andrey Karsakov[1], Alexey Yakovlev[1,2], and Sergey Kovalchuk[1]

[1] ITMO University, Saint Petersburg, Russia
[2] Almazov National Medical Research Centre, Saint Petersburg, Russia
olegmetsker@gmail.com, kesarevs@gmail.com,
ekaterina_bolgova@corp.ifmo.ru, golubev1251@corp.ifmo.ru,
karsakov@corp.ifmo.ru, yakovlev_an@almazovcentre.ru,
kovalchuk@corp.ifmo.ru

**Abstract.** This article describes the results of multidisciplinary research in the areas of analysis and modeling of complex processes of treatment on the example of patients with cardiovascular diseases. The aim of this study is to develop tools and methods for the analysis of highly variable processes. In the course of the study, methods and algorithms for processing large volumes of various and semi-structured series data of medical information systems were developed. Moreover, the method for predicting treatment events has been developed. Treatment graph and algorithms of community detection and machine learning method are applied. The use of graphs and machine learning methods has expanded the capabilities of process mining for a better understanding of the complex process of medical care. Moreover, the algorithms for parallel computing using CUDA for graph calculation is developed. The improved methods and algorithms are considered in the corresponding developed visualization tool for complex treatment processes analysis.

**Keywords:** Graph mining, Process mining, Community detection, Process modeling, Cardiology, Complex process analysis

## 1 Introduction

Data modeling is traditionally the way to understand better the processes, which provides excellent opportunities to foresee changes. The data of medical information systems, describing the processes of health care, contains empirical information about the treatment of patients, on the basis of which it is possible to develop models of these processes [1]. Electronic medical records (EHR) describing the process of providing medical care consist of many different types of process elements. Information in the EHR is contained both in a structured form and semi-structured form (for example, protocols of operations, descriptions of diagnoses, anamnesis of the patient's diseases, diaries of observation in resuscitation, conclusions about examinations, and other medical records).

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22750-0_65

The aim of this study is to develop tools and methods for the analysis of complex processes of medical care based on poorly structured data of medical information systems using high-performance algorithms. During solving this scientific problem, methods of adaptation of process mining technology have been developed to identify a complex process of providing medical care, taking into account the personal characteristics of the patient. Appointment of treatment procedures with the individual characteristics of the patient in mind contributes to modern approaches to the organization of health care (value-based health care, P4 medicine). The presence of predictive models based on empirical information is a qualitative characteristic for the health system in the management of the quality of treatment.

## 2 Conceptual Approaches to Process Space Analysis

The health care processes for different categories of patients may vary significantly. For patients in the same homogeneous group, the providing medical care process can be practically typical. However, in complex cases with concomitant pathological processes on the results of treatment of patients with cardiovascular disease can affect a significant number of factors.

### 2.1 Phase Space Analysis and Variability Reduction

The first stage of development of the patient-treatment model: reducing the variability of the phase space of the rendering process of patients with cardiovascular diseases. Process mining methods use discrete data about events of treatment episodes to identify typical ways of providing medical care, patterns identification, and deviations from the standard ways. For presentation and primary analysis of the processes of medical care in the field of the basic relational structure of the undirected graph is used, namely the identification of community structure, relationships, and relationships, or abstracted links between the instances of the processes. Also, it is possible to quantify the network function of the spread of diseases, based on random local interactions [2]. At this stage, the model is often determined as a complete systematized graph or a random graph as Erdos-Renyi or Poisson graphs [3], [4]. These models are used to quantify network characteristics such as connectivity, the existence, and size of a giant component, the distribution, and extent of elements, the degree of connectivity, and the determination of node parameters, including the clustering factor. In graphical models, the data are considered as a set of random variables, indexed nodes of the graph, where probabilistic dependences between elements are fixed. For example, directed graphs [5] [2] in the form of Bayesian networks, where each random variable is independent of others. Undirected graphical models, also called Markov random fields [6], [7], describe processes, where variables defined on two sets of nodes, are statistically independent. A key tool in working with graphical models is the Hammersley–Clifford theorem [6], [8], [9], with the corresponding positivity of the conditions, the factors of the joint distribution of the graphical model are equal to the product of potentials.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22750-0_65

### 2.2 Graph-Based Process Space Representation

It is proposed to use the graph representation of space (GPS) process model states $M\colon G^{(M)} = \langle V^{(M)}, E^{(M)} \rangle$. GPS can be used to study the functional and operational characteristics of the existing base of precedents. Each vertex of the graph corresponds to a subset of admissible model realizations of the object of research, which are considered to be identical within the framework of this interpretation (in the limiting case, the subset consists of a single instance): $v_i^{(M)} \in V^{(M)}\colon v_i^{(M)} = \left\{ v_{i,j}^{(M)} \right\} \subset M, \forall v_{i,j_1}^{(M)}, v_{i,j_2}^{(M)} \in v_i^{(M)}\colon \delta\left(v_{i,j_1}^{(M)}, v_{i,j_2}^{(M)}\right) = 0$, where $\delta$ – a specified proximity measure, usually defined as $\delta\colon M \times M \to \mathbb{R}$. The edges of the graph correspond to the level of proximity not lower than some boundary $e = \langle v_{i_1}^{(M)}, v_{i_2}^{(M)} \rangle \in E^{(M)}\colon \forall v_{i_1,j_1}^{(M)} \in v_{i_1}^{(M)}, v_{i_2,j_2}^{(M)} \in v_{i_2}^{(M)}\colon \delta\left(v_{i_1,j_1}^{(M)}, v_{i_2,j_2}^{(M)}\right) < \delta_0$ and can have weights varying depending on the actual proximity of the elements in the composition of the vertices according to a given measure $\delta$. The topology and properties of the graph allow us to analyze the features of the functional and operational characteristics specified in the space $M$: to identify the cluster, to simplify the assessment of the proximity measures for the partially observed instances of the studied objects, to assess possible alternatives to the development of situations, etc. As a result, the most significant effect of this approach is observed within the framework of the Big Data concept and the corresponding models.

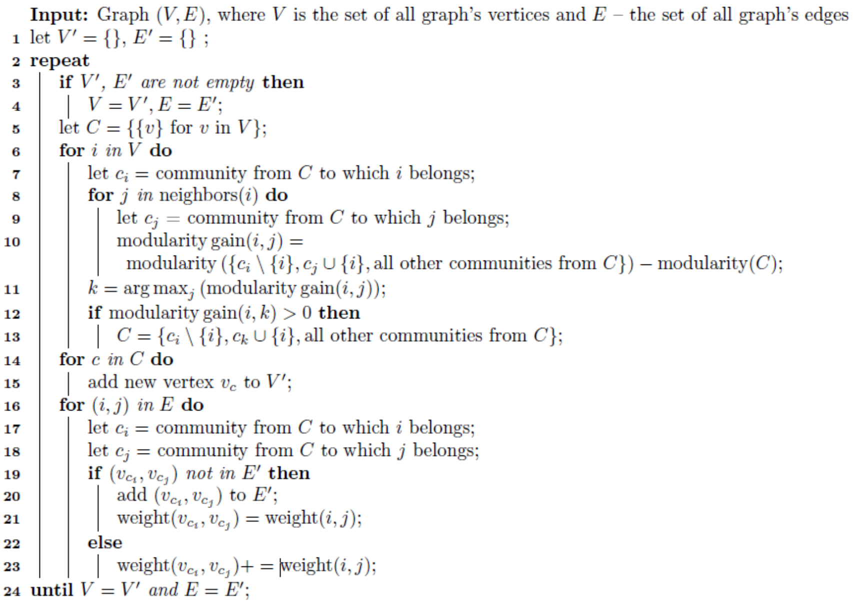
At the first stage of the study, the journal of events on the treatment of cardiovascular planned patients with stenting and ACS (acute coronary syndrome) patients (12900 patients) of intensive cardiac care unit (ICCU) and cardiac departments were analyzed using traditional process mining tools[1]. The results were poorly interpreted because the process map was characterized by variability and high complexity. As a result, we developed our solution for visualization of the phase space of the treatment process.

The edges of the graph denote the value of the symmetric difference $(\delta\left(v_1^{(M)}, v_2^{(M)}\right) = v_1^{(M)} \Delta v_2^{(M)})$ between ordered sets, the vertices denote the sets of processed events.

The primary purpose of this phase of the experiment is automatic detection the communities of the process graph. Graph describing the process of medical care and the differences in care for different groups of patients with cardiovascular diseases in the intensive care unit was developed. Community detection is a way to highlight the structure of the phase space. The algorithm of community detection is proposed in figure 1.

The quality of partitions proposed by the community detection algorithm is estimated by the metric called modularity. The algorithm proposed for iterative optimization of the modularity score [10]. It is designed to work with weighted undirected graph structures.

[1] http://www.fluxicon.com/disco/



```
Input: Graph $(V, E)$, where $V$ is the set of all graph's vertices and $E$ – the set of all graph's edges
let $V' = \{\}$, $E' = \{\}$ ;
repeat
    if $V'$, $E'$ are not empty then
        $V = V', E = E'$;
    let $C = \{\{v\}$ for $v$ in $V\}$;
    for $i$ in $V$ do
        let $c_i$ = community from $C$ to which $i$ belongs;
        for $j$ in neighbors$(i)$ do
            let $c_j$ = community from $C$ to which $j$ belongs;
            modularity gain$(i, j)$ =
             modularity $(\{c_i \setminus \{i\}, c_j \cup \{i\},$ all other communities from $C\})$ – modularity$(C)$;
        $k = \arg\max_j$ (modularity gain$(i, j)$);
        if modularity gain$(i, k) > 0$ then
            $C = \{c_i \setminus \{i\}, c_k \cup \{i\},$ all other communities from $C\}$;
    for $c$ in $C$ do
        add new vertex $v_c$ to $V'$;
    for $(i, j)$ in $E$ do
        let $c_i$ = community from $C$ to which $i$ belongs;
        let $c_j$ = community from $C$ to which $j$ belongs;
        if $(v_{c_i}, v_{c_j})$ not in $E'$ then
            add $(v_{c_i}, v_{c_j})$ to $E'$;
            weight$(v_{c_i}, v_{c_j})$ = weight$(i, j)$;
        else
            weight$(v_{c_i}, v_{c_j})$+ = weight$(i, j)$;
until $V = V'$ and $E = E'$;
```

**Fig. 1.** The algorithm of community detection of close cases of treatment

### 2.3 Communities Interpretation and Modelling

The second stage of development of the patient-treatment model: analysis of the treatment process of patients in particular groups. In the second stage, it is possible to analyze and simulate processes within each individual community. At this scale of treatment process detailing it is possible to calculate the probability of events and treatment pathways with the least variability. For example, the most frequent event of patients with the hypertensive disease is "primary examination consultation of a cardiologist" (61,507 times in 40,537 episodes). Most patients who have passed the necessary examinations at the stage of applying for qualified help or at the primary level are sent to the initial consultation of a specialist. Some of the patients who applied need echocardiography, if this study was not carried out, more than six months have passed since the previous study or a protocol of low-quality study is presented.

Graphs are a more general class of structures than sets, sequences, and data trees. Graph mining is used to analyze repetitive patterns and perform specificity, discrimination, classification, and cluster analysis of large data sets. Using medical data, it is possible a specification of the clusters of the graph events of treatment for patient data in the cluster (community). Classification and cluster analysis of graph datasets can be studied by integrating them with the process of identifying frequent graph patterns. Analysis of similarity measures of events sets to analyze the nodes proximity of the graph of treatment of patients with cardiovascular diseases can be used to predict and rank new treatment processes, and the study of empirical data already accumulated in the medical information system (MIS).

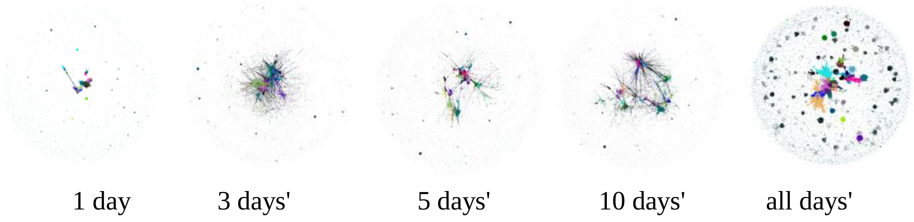
Most methods of data mining of graph clusters suggest that the results do not depend on the personal characteristics of the patient. There may be relationships between individual characteristics and patients treatment graph clusters (communities). Identification of links between objects in such networks provides an assessment of the significance of different links concerning data on the treatment process and the patient, which may be the basis for requesting this information in the next stages of treatment. For the clinical process, this can quantitatively justify the priority of diagnostic procedures. In terms of treatment quality - to define quality indicators and priority of their implementation.

## 3 The Model Development

Treatment processes are directly related to the processes occurring in the human body and vice versa. The analysis of complex treatment processes in general involves a system of two main types of models: patient models (models of pathophysiological processes, calculation of the clinical events probability, including the probability of death and the risks of complications, etc.); models of medical care processes (treatment processes, administrative processes, logistics processes, the effects of therapy, the appropriateness of the procedure).

### 3.1 Prediction of Processes' Dynamics

In the next experiment on the analysis of data from the MIS determined the content and structure describing the treatment process in the form of a graph. Data on the treatment process are discrete and represent a variety of events (for example, admission to the department, echocardiography, heart surgery, transfer to another department, general blood test, etc.). The symmetric difference between ordered sets of patient treatment events is a method of calculating the weight of the graph edge, expressed in the number of different elements. The algorithm identified communities described above. Further calculations for several graphs at various stages of treatment are made: one day, three days, five days, ten days, all events (Fig. 2)



1 day 3 days' 5 days' 10 days' all days'

**Fig. 2.** Representation of the phase space of treatment processes at different stages.

It was possible to identify patients in terms of movement in these communities find similar trajectories and classify them by machine learning methods for trajectory prediction.

In the obtained graphs, some vertex communities are founded, which at different stages of treatment have different compositions (eczema-pliers of patient treatment processes). It was possible to identify the pattern of treatment processes in terms of movement in these communities at various stages using machine learning methods to predict the path in the areas of phase space.

Then the problem of classification into a certain class of patients is solved. Machine learning methods (SVC, Random Forest, KNeigbor, Logistic Regression, Naïve bayes, GB - gradient boosting, ensemble V1 = RandomForest + KNeigbor, ensemble V2 – RandomForest + LogisticRegression) were used to train the model. The following predictors were used: minimum hemoglobin level, maximum troponin, maximum ALT, maximum AST, maximum creatinine level, maximum PLT level, maximum glucose level, age, gender). Ensemble V2 showed coating ROC 88 %.

## 4 Experimental Setting with GraphMiner Toolbox

Interactive toolbox developed for the analysis of complex processes taking into account all the methods described above. In this study, we conduct an experiment using this tool to analyze complex patient-treatment processes. The main elements of the software package:

- Knowledge base with data analysis (Data Mining/Process Mining/Text Mining).
- Computing core. The experimental study calculations were implemented using the GPU.
- Interactive Visualizer [11]. Supported processor architecture: x86, x86-64, supported platform: .NET 4.0 Programming language: C#

The visualization of a resulting graph with a force-directed layout algorithm allows the user to analyze patients' communities further, find misbehaviors in an algorithm and detect a different clustering or higher-level similarities between the communities. The overview of the visualization tool is demonstrated in the video available on YouTube[2]. This tool visualizes a three-dimensional non-oriented force-directed graph layout, where nodes are patients and links between them depending on their likelihood, calculated in a previous chapter. The user interface allows the user to further explore and analyze the graph with different actions.

[2] https://youtu.be/EH74f1w6EeY

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22750-0_65

## 5 Conclusion and Future Work

The relevance of the development of custom tools of intellectual analysis, taking into account the specifics of the subject area of data and the processes described by them is not in doubt. This practice integrates domain-specific knowledge with data analysis techniques and provides data mining solutions for specific areas. Graph mining visualization integrates visual elements into data mining, process mining to discover implicit and useful knowledge from large sets of medical data. Parallel computing technologies provide a reasonable response in the calculations and interactions. For the analysis, MIS data offered several well-proven methods: classification, clustering, analysis of the significance of predictors, correlation analysis, etc. Trends in data mining and the methods considered in this study (improved scalable integration of data mining with data storage and interactive knowledge bases).

Further work is possible in the direction of studying various graphs. The issues of identifying the community, improving the performance of calculations, machine learning methods, visualization of results are also relevant in further research. It is also possible to scale these methods and toolbox for other process areas.

**Acknowledgments.** This work financially supported by the Ministry of Education and Science of the Russian Federation, Agreement #14.575.21.0161 (26/09/2017). Unique Identification RFMEFI57517X0161.

## References

1. Yang L., Zhang J. Automatic transfer learning for short text mining // EURASIP J. Wirel. Commun. Netw. 2017. Vol. 2017, № 1. P. 42.
2. Chamley C. Rational Herds: Economic Models of Social Learning. Cambridge Univ. Press, 2004.
3. M. Jackson. Social and Economic Networks // Princet. Univ. 2008.
4. Newman M. Networks: An Introduction // Oxford Univ. Press. 2010.
5. Edwards D. Introduction to Graphical Modelling // Springer. 2000.
6. Kindermann and J. L. Snell. Markov Random Fields and Their Applications // Am. Math. Soc. 1980. Vol. 12.
7. Willsky A.S. Multiresolution Markov Models for Signal and Image Processing. 2002. Vol. 90, № 8.
8. Besag J. Spatial interaction and the statistical analysis of lattice systems. J. Royal Stat. Soc., 1974. P. 192–236, 1974.
9. Handscomb J.M.H. and D.C. Monte Carlo Methods // Chapman Hall. 1964.
10. Blondel V. et al. Fast unfolding of communities in large networks // iopscience.iop.org.
11. Karsakov A. et al. Toolbox for Visual Explorative Analysis of Complex Temporal Multiscale Contact Networks Dynamics in Healthcare // core.ac.uk.

DOI: 10.1007/978-3-030-22750-0_65