



**I T H E A**



**International Journal**

**INFORMATION THEORIES  
&  
APPLICATIONS**



**2016 Volume 23 Number 1**



International Journal  
INFORMATION THEORIES & APPLICATIONS  
Volume 23 / 2016, Number 1

Editorial board

Editor in chief: **Krassimir Markov** (Bulgaria)

<b>Alberto Arteta</b>	(Spain)	<b>Luis F. de Mingo</b>	(Spain)
<b>Aleksey Voloshin</b>	(Ukraine)	<b>Lyudmila Lyadova</b>	(Russia)
<b>Alexander Eremeev</b>	(Russia)	<b>Martin P. Mintchev</b>	(Canada)
<b>Alexander Kleshchev</b>	(Russia)	<b>Natalia Bilous</b>	(Ukraine)
<b>Alexander Palagin</b>	(Ukraine)	<b>Natalia Pankratova</b>	(Ukraine)
<b>Alfredo Milani</b>	(Italy)	<b>Rumyana Kirkova</b>	(Bulgaria)
<b>Avram Eskenazi</b>	(Bulgaria)	<b>Stoyan Poryazov</b>	(Bulgaria)
<b>Boris Fedunov</b>	(Russia)	<b>Tatyana Gavrilova</b>	(Russia)
<b>Constantine Gaidric</b>	(Moldavia)	<b>Tea Munjishvili</b>	(Georgia)
<b>Galina Rybina</b>	(Russia)	<b>Valeriya Gribova</b>	(Russia)
<b>Hasmik Sahakyan</b>	(Armenia)	<b>Vasil Sgurev</b>	(Bulgaria)
<b>Ilia Mitov</b>	(Bulgaria)	<b>Vitalii Velychko</b>	(Ukraine)
<b>Juan Castellanos</b>	(Spain)	<b>Vitaliy Lozovskiy</b>	(Ukraine)
<b>Koen Vanhoof</b>	(Belgium)	<b>Vladimir Donchenko</b>	(Ukraine)
<b>Krassimira B. Ivanova</b>	(Bulgaria)	<b>Vladimir Jotsov</b>	(Bulgaria)
<b>Leonid Hulianytskyi</b>	(Ukraine)	<b>Vladimir Ryazanov</b>	(Russia)
<b>Levon Aslanyan</b>	(Armenia)	<b>Yevgeniy Bodyanskiy</b>	(Ukraine)

International Journal "INFORMATION THEORIES & APPLICATIONS" (IJ ITA)  
is official publisher of the scientific papers of the members of  
the ITHEA International Scientific Society

IJ ITA welcomes scientific papers connected with any information theory or its application.

IJ ITA rules for preparing the manuscripts are compulsory.

The **rules for the papers** for IJ ITA are given on [www.ithea.org](http://www.ithea.org).

Responsibility for papers published in IJ ITA belongs to authors.

International Journal "INFORMATION THEORIES & APPLICATIONS" Vol. 23, Number 1, 2016

Edited by the Institute of Information Theories and Applications FOI ITHEA, Bulgaria, in collaboration with:

Institute of Mathematics and Informatics, BAS, Bulgaria,

V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,

Universidad Politécnica de Madrid, Spain,

Hasselt University, Belgium,

St. Petersburg Institute of Informatics, RAS, Russia,

Institute for Informatics and Automation Problems, NAS of the Republic of Armenia

Printed in Bulgaria

Publisher ITHEA®

Sofia, 1000, P.O.B. 775, Bulgaria. [www.ithea.org](http://www.ithea.org), e-mail: [info@foibg.com](mailto:info@foibg.com)

Technical editor: Ina Markova

Copyright © 2016 All rights reserved for the publisher and all authors.

© 1993-2016 "Information Theories and Applications" is a trademark of ITHEA®

© ITHEA is a registered trade mark of FOI-Commerce Co.

ISSN 1310-0513 (printed)

ISSN 1313-0463 (online)

## IMPLEMENTING A LINEAR FUNCTION TO MEASURE THE QUALITY IN GOVERNMENTS

Alberto Arteta, Juan Castellanos, Yanjun Zhao, Danush K. Wijekularathna

**Abstract:** *Once the biggest issue with the current democracy system is identified, the research focuses on finding the root of the problem. This is a key step and requires further analysis as we are constantly deviating from the true origin. Unless we tackle the root of the problem any attempt we make to improve the system will be in vain. The current democratic system is severely flawed and therefore requires improvement that can only be detected by society. The works highlighted here report numerous problems of corruption, lack of transparency, information monopoly, and bipartisanship in virtually all current governments. These days, citizen demonstrations against political abuse are a social reality, which proves the large gap that exists between the government and the middle class.*

*This work defines several parameters to measure the quality of governments and propose to use a linear function to check on them.*

*Once the root of the problem comes to the surface, the way to move forward becomes clear.*

**Keywords:** *Linear function, quality measurement, Democratic system*

---

### Introduction

---

By accepting that change is needed because that is what we see in the streets very often, the change that everyone wants is a government that preserves the social welfare by creating jobs, boosting the economy and protects the weak ones. The aspirations of change are focused on the choice of good government. Therefore, a very interesting and necessary point in the election process is to define which government is a good one. Several authors have written about the qualities that a good government should have. John Gant calls on transparency as a major requirement. Charles Murray a senior researcher of policy research institute Manhattan in his book *In Pursuit: Of Happiness and Good Government* [Raggat, 2007] advocates for a more decentralized government. Furthermore, government corruption is objects of numerous books. Obviously, it is desirable to reach a zero level of corruption in public positions.

In the opinion of most experts, and particularly the governments, the goodness or the quality of a government is linked to the ability of making "right" decisions [Murray,1994], [BrainBridge,2008], meaning that the decision is motivated by the public interest and ultimately favours this. What is clear is

that a government cannot do anything. It is not merely decorative [Munshi, 2004]. Otherwise, we could remove it and save a bunch of money. So as It must make decisions, this decisions must be the best ones or the right ones, meaning that they have to help to achieve preserving or improving security and social and economic welfare for the citizens they represent, (or at least minimize the effects of an unavoidable damage to the country).

---

### First approach

---

Some political scientists establish that there are mainly three factors accepted as valid to measure the quality of a government. These are vocation, preparation and competence

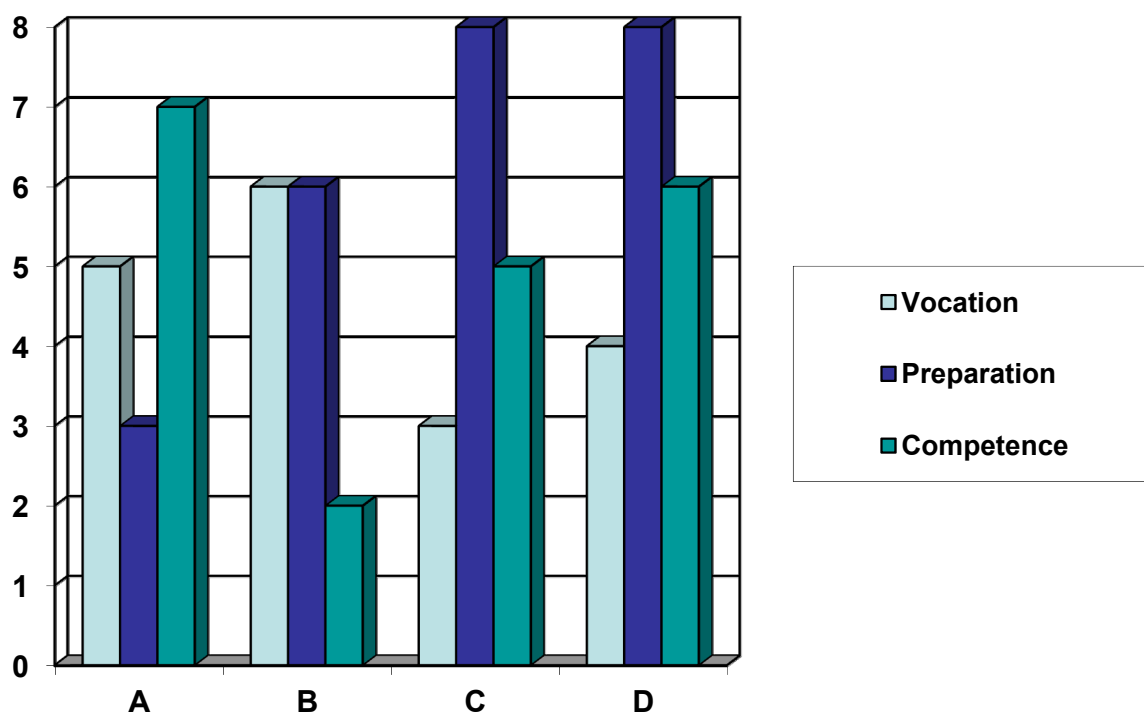


Figure 1. Vocation, Preparation and Competence

This would be an example of measuring 4 governments (A, B, C, D) that meet the criteria of vocation, competence and preparation.

The analysis carried out below is open to debate and it's not scientifically proven. It is a proposal based on political analysts' opinions and observable facts. However I consider it can provide a fair method to evaluate a government performance.

In order to assess the quality of a government more factors could be considered, but obviously these three are essential. In this proposal, the same weight is assigned to each of these factors which mean that we accept they are equally important to evaluate the government's quality.

---

### Vocation

---

Peter Raggat [ Raggat, 2007] is a writer and expert on the educational systems of UK. In his book "**Government, Markets and Vocational Qualifications: An Anatomy of Policy**" Raggat conducts a review of decision policies. In his opinion, there is a vocational education program in the UK, however this does not necessarily empower growth and welfare. According to Raggat, the political class is the one who profit the most from the government decisions. Raggat defines vocation as a fundamental value. Logically, vocation and corruption are two opposite concepts to measure the degree of involvement in the state tasks. The more vocation exists in the performance of the tasks (Serving your country), the least corruption will be. We understand that on a scale (0.10), 10 is absolute vocation meaning that only the general interest prevails in the country and there is no personal interest at all in the political duty. 0 would be absolute corruption meaning there is no vocational interest at all, the only motto then is private interests (personal enrichment). Any intermediate value will have some vocational and corrupt part (which is probably where most of the existing governments are).

The correct decision-making is undoubtedly influenced by the vocational level. A 100% vocational decision will always seek the general interest of the citizens. A decision that has absolutely no sense of vocation but only the personal interest, will be 100% corrupt. There will be decisions that maybe are good for public interest and also for the individual. However, the main motivation that should drive decisions of a government should be the public interest. The more vocational, less corrupt government decisions are (regardless they are wrong or right). Therefore that should be the first step to be taken by the government that aspires to have the highest quality.

---

### Preparation

---

Again, if we admit that society has to be prepared to carry out whatever tasks they perform, it is not lower the need for elected leaders to be prepared, trained and fully qualified to rule and serve their country during the period they are in power [Brifault,2000]. The preparation may come from academic studies, courses, conferences, and experience. Obviously someone who has ruled a country has gained experience in decision-making and it is normal that they have learned from that experience. This logically adds a good value when decisions need to be made. So, when a government is experiencing a situation that requires making a decision, they should know what to do. Obviously, they must be prepared and know the consequences of their decisions[Caplan,2007]. This preparation clearly should have come from their experience and training. This proposal considers that preparation (*prep*) consist of

two equally important factors will  $prep = \frac{1}{2}exp + \frac{1}{2}training$  . "exp" is variable that returns the experience and "training" the theoretical knowledge

---

### Competence

---

This is possibly the differential factor 3 because surely is what makes the difference between all governments. This is defined as the government's ability to solve problems, make the right decisions in difficult times. It is the point of genius that should be distinguished from the other. It's that ability that shows that capacity of reactions facing unpredicted and unforeseen situations, the response must be the best as possible. It is the innate ability that develops in some people, usually driven by a vocation when a situation requires to be solved without the manual. This factor is noticeable in many examples such as the surgeon who innovates a method that was not on the books to operate and heal the patient when this was about to pass away, the teacher who gets the best out of the student without following general guidelines, that ability to do your job well, not always by following the established recipes. Naturally, this is impossible without vocation or preparation.

A person who has no medical training or vocation, although it has a potential of making good decisions, may not be as competent as a doctor who does have those two values.

Only with the vocation we might not be able to be competent because we would need to have the knowledge on which to base our decisions. And only with the preparation, we will be able to apply the recipes in the book but it would be very difficult to have that ability to do our job successfully if we are not driven by the call to serve our citizens. We might lose in situations that are not in our recipe guide.

The interesting thing is that while the vocation and preparation are essential to be competent, their existence does not always guarantee a high value of competence. And that is the differentiating factor.

Sure there will be many doctors minded and prepared but with little power to make decisions when they face unexpected situations that are beyond the script. So, although preparation and dedication are essential to be competent, it takes a touch of genius, that ability that really makes the difference between a good government and the optimal one [Chomsky, 2002].

---

### Building the function

---

Our analysis considers the three values equally important (vocation, preparation, competence). We can then establish the following estimation for the quality of governance:

$$F(x, y, z) = \frac{1}{3}x + \frac{1}{3}y + \frac{1}{3}z$$

x=vocation, y=preparation z=competence

As mentioned, the competence will depend largely on the previous two, so it can be defined as a function with variables: Vocation and preparation influences the ability to make decisions when unexpected situations arise.

In addition it will be needed to take into account the relation between corruption factor and time. The level of corruption in a government can increase upon time. The longest a government stay in power the higher are the chances for the corruption to increase. So, although experience adds value to the preparation, it is also clear that might increases the level of corruption in long term. In fact, there are many cases of governments that have lasted more than 20 years and have proven a high level of known corruption cases.

This analysis suggest that government acquires a higher preparation level the longer they stay power but its vocational level tends to decrease over time by increasing the corruption. We understand that the relation between corruption and time is a generalized linear progression for some societies, being possibly quadratic or exponential in underdeveloped societies. (Probably related to the training level of the population) Following, there is a model example for the progression of experience and corruption versus time. These would be the functions  $\text{exp}(t)$  and  $\text{voc}(t)$ , which measure vocational level and experience of a particular government "A" upon time.

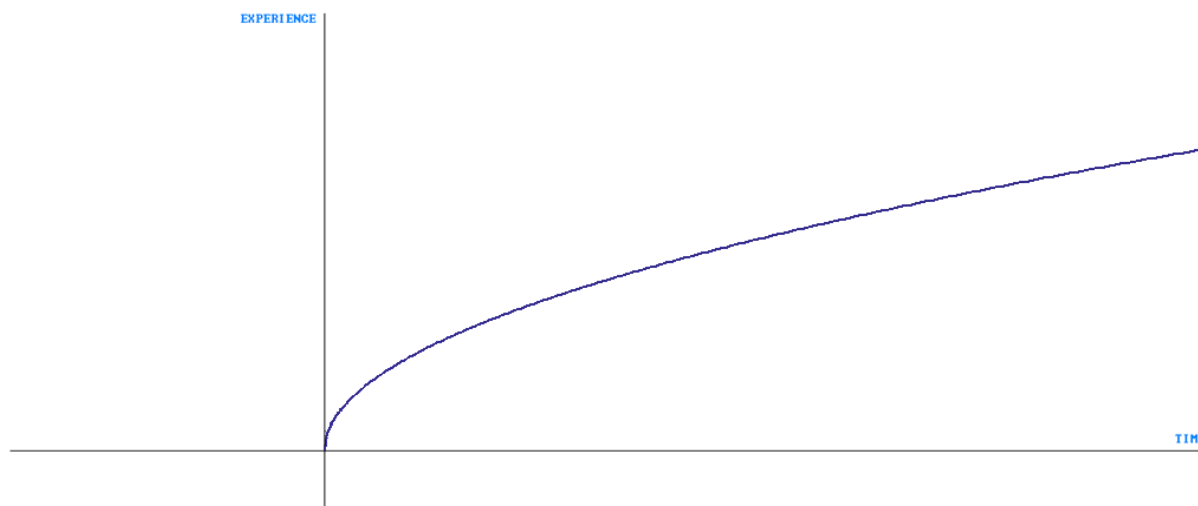


Figure 2. Function  $\text{Exp}(t)$  experience increases with time (t) years in office. Government Experience(A)

Obviously the experience has a limit and that is why the representation has to show a curve. It is impossible to have infinite experience

However, it seems logical to think that corruption also increase upon time. The proof of this is governments who have spent many years in power and have degenerated into extreme corruption

diminishing their vocational level. The graph shows how the vocation decreases smoothly as we move through time in a standard government. Depending on the situation, perhaps the vocation could even increase over the early years and then decrease in the long term, or maybe vocation decreases faster, but in general the governments that have been long in power lose vocation and gain in corruption.

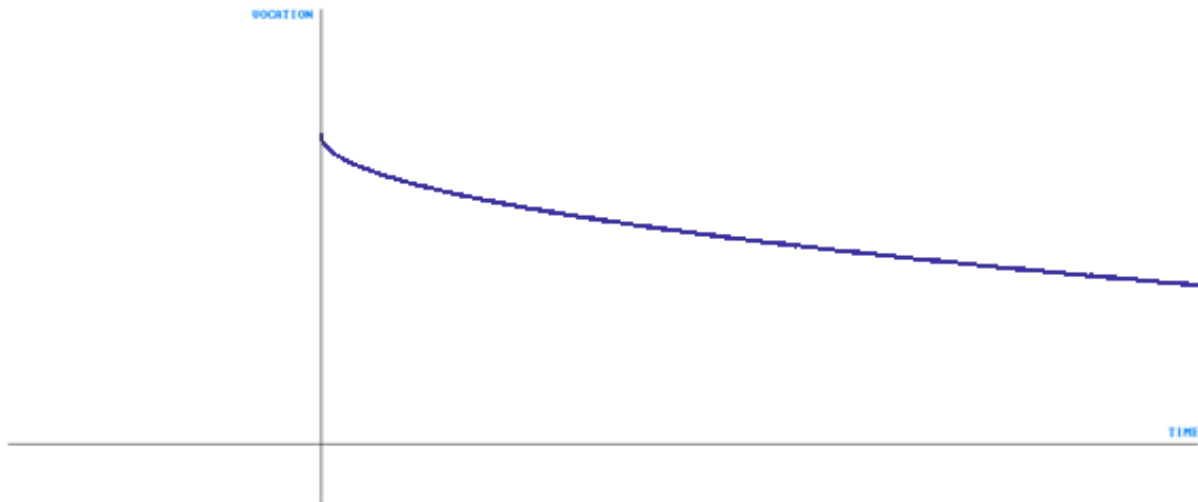


Figure 3. Corruption increases the longer in power, and therefore vocational level declines gradually  
 $VOC(t)$

Finally the level of competence of a government  $COMP(x, y, s)$  can univocally be defined, as competition depends on the vocation ( $x$ ), preparation ( $y$ ) and the innate ability of making correct decisions ( $s$ )

In this proposal these facts are accepted:

- Competence is 0 without any vocation.
- Competence is 0 without any preparation
- Competence is 0 without any innate ability to face unexpected problems.

We also accept that these three values equally influence the competence factor existing in any government.

So  $COMP$  can be defined this way:

$$COMP(x, y, s) = \begin{cases} 0 & x = 0 \vee y = 0 \vee s = 0 \\ \frac{1}{3}x + \frac{1}{3}y + \frac{1}{3}s & x \neq 0 \wedge y \neq 0 \wedge s \neq 0 \end{cases}$$



The function that combines COMP, VOC, PREP could be a good indicator for the quality of a government.

For example, let us suppose a government 'A' with people who have been long in power, a great political party; a very small privileges or almost nil-paid. There are no known corruptions cases. Moreover, we have seen that its policies have helped the economy and there have been a social improvement. Furthermore, it has proven high professionalism when dealing with difficult issues. This government would get a high grade.

However, we can evaluate a government 'B' whose members have profited, with no political preparation and just have not shown that their decisions have helped citizens but impoverished them. There have been cases where governments have much higher level of private and public deficit without economic growth lead. In that case it's the grade could be quite low.

So, below there is a proposed method for assessing the quality of any of the existing government:

We see the time they have been in office:

Based on that, experience, level of political education: (qualifications, training, degrees), known corruption cases, benefits and privileges granted to government officials.

With this info two parameters are fixed:

- Preparation: (experience and training)
- Vocation (time in office, corruption)

The Innate ability to solve problems by creating policies that have proven effective has an influence in the COMP factor.

With the top 2 parameters ("preparation", "Vocation") are found; from them would be possible to infer the third parameter "competence".

Eventually F (COMP, VOC, PREP) would return the index of quality of government. To maximize the value of quality, COMP values, VOC, PREP must also be maximized. However, as we have defined it, to greatly increase the level of experience, vocation decreases.

The function F reaches maximum values when governments are highly trained, have several years of experience, present no corruption or have obtained benefits from office, and have demonstrated that their actions have benefited the country they represent.

Among the first government prepared 'A' and the second 'B', there may be many intermediate. It seems logical that these functions could be a good starting point for assessing the management of whatever government.

---

### Conclusions and future work

---

The need to have high quality systems is out of question. There are many parameters to measure how good a system is. This work proposes a method to measure the quality of a democratic system through the use a function that returns a value. Most ways to determine the quality of governments tend to be subjective; however this method combines 3 objective values to have the most possible objective value when dealing with measuring that quality. This function gets numeric values that match the competence, vocation and preparation. Therefore these values represent a reliable way to establish the quality of a specific government.

The application of this function to the current governments might show that the quality of the current governments leave a big room for improvement. Thus, this is a first step to start a new revolutionary way to try to improve the quality for whatever system is object of study and more specifically the democratic systems.

---

### Bibliography

---

- [BrainBridge, 2008] The New Corporate Governance in Theory and Practice, Author: Stephen Bainbridge. Publication Date: July 23, 2008 | ISBN-10: 0195337506 | ISBN-13: 978-0195337501
- [Brifault, 2000] Dollars and Democracy: A Blueprint for Campaign Finance Reform ,Author: Richard Briffault, Association of the Bar of the City of New York (Corporate Author) Publisher: Fordham University Press; 2 edition (January 1, 2000, ISBN-10: 0823220966 Publication Date: January 1, 2000
- [Caplan, 2007] The Myth of the Rational Voter: Why Democracies Choose Bad Policies, Author: Bryan Caplan, Publisher: Princeton University Press (April 16, 2007) ISBN-10: 0691129428, ISBN-13: 978-0691129426
- [Chomsky, 2002] Media Control, Second Edition: The Spectacular Achievements of Propaganda (Open Media Series) , 64 pages Publisher: Seven Stories Press; 2 Sub edition (September 3, 2002) ISBN-10: 1583225366, ISBN-13: 978-1583225363
- [Munshi, 2004] Good Governance, Democratic Societies and Globalization, Author: S. Munshi, Paul Abraham, Sage Publication, 2004 , 368 pages ISBN: 9780761998488
- [Murray, 1994] In Pursuit : Of Happiness and Good Government, Author: Charles Murray , 300 pages, Publisher: ICS Press Mayo 1994, ISBN-10: 1558152970 ' ISBN-13: 978-1558152977
- [Raggat, 2007] Government, Markets and Vocational Qualifications: An Anatomy of Policy Authors: Steve Williams, Raggat ISBN: 0750709162, Editor: Taylor & Francis 16 de marzo de 2007

**Authors' Information**

---



**Alberto Arteta Albert** – Computer Science Dpt, Assistant Professor, Troy University  
Wright Hall 112E University Avenue Troy, 36081 AL USA email: [aarteta@troy.edu](mailto:aarteta@troy.edu)



**Juan Castellanos** – Artificial Intelligence DPT Faculty of Informatics Universidad  
Politécnica de Madrid (UPM); Campus Montegancedo, UPM; e-  
mail: [jcastellanos@fi.upm.es](mailto:jcastellanos@fi.upm.es)



**Yanjun Zhao** - Computer Science Dpt, Assistant Professor, Troy University MSCX 129B  
University Avenue Troy, 36082 AL USA email: [yjzhao@troy.edu](mailto:yjzhao@troy.edu)



**Danush K Wijekularathna** Department of Mathematics, Assistant Professor, Troy  
University, Wright Hall 112B University Avenue Troy, 36081 AL USA  
Email: [dwijekularathna@troy.edu](mailto:dwijekularathna@troy.edu)

## COLLECTIVE COMPUTATION: TURNING THE UNDERGROUND INTO AN ANT NEST

Clemencio Morales, Luis Fernando de Mingo

**Abstract:** *The management and proper use of the Urban Public Transport Systems (UPTS) constitute a field as critical as little investigated according to its relevance and urgent idiosyncrasy within smart cities realm. In this paper, a newfangled approach by using the Natural Computing paradigm and Collective Computation is shown, more concretely taking advantage of an Ant Colony Optimization algorithm variation in order to build a system that makes the complete control of the UPTS a tangible reality.*

**Keywords:** *Smart City, Natural Computing, Collective Computation, Urban Public Transport System, Wireless Sensor Networks*

**ACM Classification Keywords:** *10003120.10003138: Human-centered computing - Ubiquitous and mobile computing, 10010147.10010178: Computing methodologies - Artificial intelligence, 10010147.10010257.10010293.10011809: Computing methodologies - Bio-inspired approaches, 10010405.10010481.10010485: Applied computing - Transportation*

**MSC:** *68Q32 Computational learning theory, 68T05 Learning and adaptive systems*

---

### Introduction

---

Since its pioneer conception at 1843, the underground trains have changed in such a mastodontic number of ways that it will take ages to enumerate them. However, these changes have not been applied to the management system and conception of the underground itself as it is nowadays. On the one hand, the rapidly-growing massification of the world’s urban cores, in collaboration with the underground intensive use by citizens, is pushing the transition of these cores to the smart city purest concept, where every single element within the city has ratiocination enough for it to be called *intelligent*. The aforementioned massification can be seen in Fig. 1.

On the other hand, it is important to note that this need has been outlined by organisms such as C.E.O.E (Spanish Confederation of Business Organizations). In fact, as described in [CEOE, 2015]:

“This frame of sustainability and efficiency that must involve the Smart Cities, has a direct relationship with other key areas, such as [...] the **efficient management of the mobility of people** [...] [Cities are lacking] Indicators for the collection appropriate measures [...] [Cities systems need] real-time knowledge about incidents, and an improved efficiency and management of the public transport”

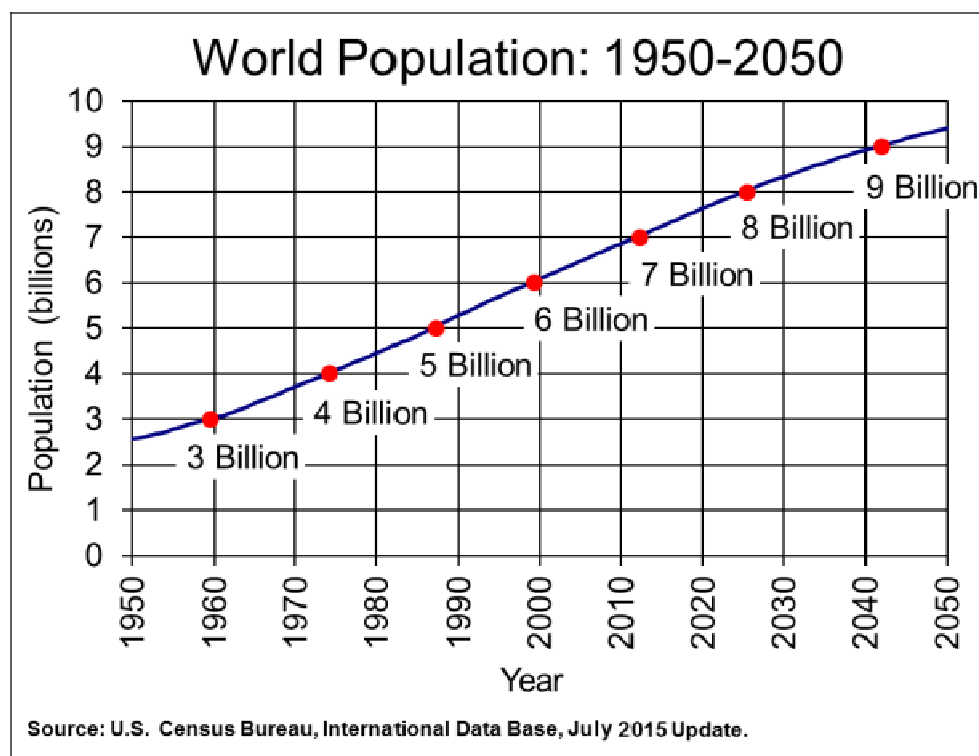


Figure 1. World Population estimated growth between the years 1950 and 2050

(Source: United States Census Bureau, International Data Base)

Even the concept of Smart City is still being under constant redefinition, most authors agree that many different individuals, agents and devices, operate with their environment within the Smart City realm. Therefore, as [Hollands, 2008] points out, the relation among all of these elements will define the behaviour of the Smart City itself. It is easy to realize that an important area of the Smart City will be based in the interaction between the different components of it with their environment. This fact disembogues in a **Socio-Collective Interaction**, where the smart city in general terms, and specially the underground system beneath, can be seen as a huge swarm, where agents collaborate between them. The aforementioned approach justifies the present project, mainly based on a change in the way of tackling the management processes of any underground system, using Collective Computation algorithms instead of the classical, graph-oriented ones.

---

### Definition of the problem

---

The underground system beneath any urban core is a living, constantly-in-change entity. According the Annual Subway Ridership of the Metropolitan Transportation Authority [MTA, 2012], 3.410 billion itineraries were made last year within Beijing underground system. Thus, how can be these itineraries traced, letting the management know who is using the underground and when? How can the users rapidly know if there is an emergency or a path which is not working due to technical errors within the

underground? How can we monetize the massive data that can be potentially generated by so many itineraries? The response to these questions is precisely the reason that justifies the ongoing project, which aims to create a synergy of elements achieved due to the application of many newfangled Computation Paradigms. These paradigms, in collaboration with a strict software of control purposes, that will operate with user's Smartphone's, will head to an increase in the intelligence of the Urban Public Transport Systems (UPTS hereinafter).

---

### Investigation goals

---

The present investigation has the objective of fixing, chiefly, the following goals, that define an accurate overview of the project investigation:

- Investigate the Computing Paradigms according to the realm of the Collective Collaboration: As it will be explained in further sections within this document, the Natural Computation stands as the best ally when it comes to this investigation aspect. This paradigm is made up by Genetic Algorithms, Ant Colony Optimization, Swarm Computing, Grammatical Evolution and Grammatical Swarm, which will be investigated in order to find possible improvements, if any.
- Find a nexus between the Computing Paradigms involved and the problem to solve: Once a strong theoretical overview has been given to the reader, the union point and nexus with the chased system will be described. It is remarkable that, at the time being, there is no application of these algorithms in the UPTS context, factor that increases the newfangled character of the present investigation.
- Design and development of a system that, using the needed paradigms within Natural Computation, allows a wide study of the behaviour of the underground users: In a nutshell, the system aims to become a tool that makes possible the study of the user's behaviour, by taking dissociated data up in order to guard the privacy of the citizens. This objective will be possible thanks to the UPTS users Smart phones, for which an application is to be developed in the Android Operating System. Please note that this system will make possible to:
  - \* Make precise studies about the statistical population that uses the UPTS.
  - \* Know, in an accurate way, the most popular routes for the users, as well as their behaviour between the UPTS. It is remarkable that this factor constitutes an open door for an efficient<sup>1</sup> management within the system.

---

<sup>1</sup> It is important to note that the term efficient differs to effective in a subtle, but crucial manner; while an effective system achieves every objective, an efficient one achieves every objective as well but in the best, optimal way.

\* Prepare, thanks to the estimations gathered from the statistical study of the data, the UPTS in order to deal with peaks. Such scenario can be predicted by attending at atypical values within the data set gathered by the system.

\* Detect anomalous situations, such a blocked train within a tunnel, or different casuistry where the number of users standing at the platform is high enough to fear an accident, surpassing the capacity of the specific dock.

\*Prepare alternative routes in case of intensive use and/or fault of the UPTS systems.

More functionalities and features, currently in an evaluation and delimitation stage.

### Topology of the system

The system to be developed formed by essentially five groups of elements, has the main structure that can be seen in Figure 2:

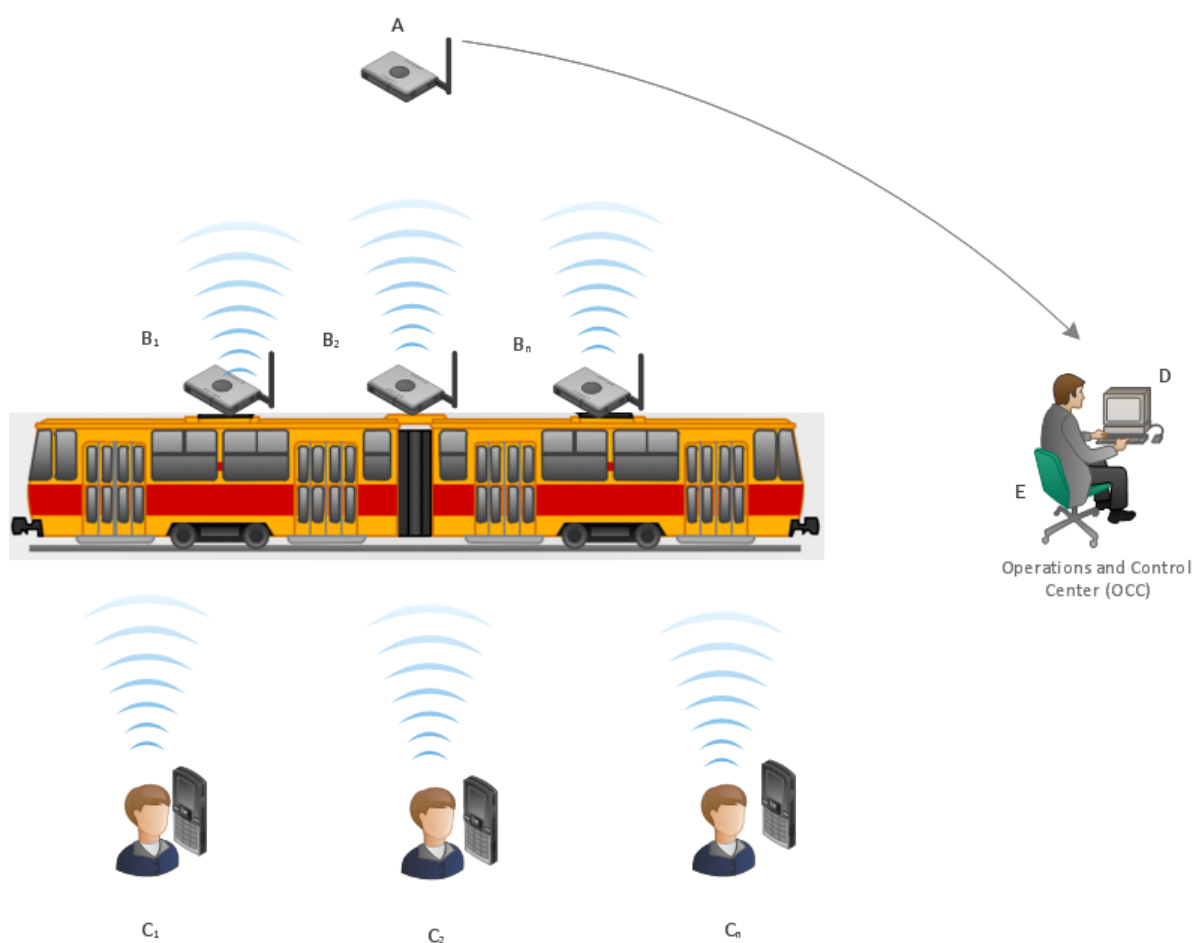


Figure 2. Overview of the currently-under-development system

As Fig. 2 shows, there are five groups of elements within the system, which are described below:

*A: Fixed Smart dust:* In Wireless Sensor Networks terms, this smart dust element, that will be unique in each UPTS station, will behave as the *sink* node. This special device will be integrated in the dock itself, receiving information from its  $B_n$  counterparts. The *sink* node will be the only element able to establish communication with the Operations and Control Center,  $D$ , and the elements within the  $C$  set.

$B[1,2,3,\dots,n]$ : *Mobile Smart dust:* The present element of the system, embedded in the UPTS trains fleet, will behave as a *slave* of the *sink* smart dust. They will establish communication with the *sink*, and will receive data from the elements within the  $C$  set, which will be explained below.

$C[1,2,3,\dots,n]$ : *Users smart phones:* This fundamental element within the system will be used tacitly by the users in order to make evident their presence in the dock. The elements within this group will be able to communicate with the *sink* smart dust,  $A$ , as well as with its counterparts in  $B$ , the *mobile smart dust* devices, that will make possible to know the number of users in the train. It is remarkable that the set formed specially by  $A$  and  $B[1,2,3,\dots,n]$  will shape the Wireless Sensor Network of the system, that will operate closely with the  $C[1,2,3,\dots,n]$  devices. A mobile application will be developed for the elements in  $C$ , that will retrieve the dissociated data of the users, let them know different routes in case of massive congestion, configure itineraries and show warning regarding abnormal situations that may occur within the UPTS.

*D: Operations and Control Center (OCC):* This element will behave as the management point within the system realm, receiving the data sent by the smart dust in every single station, showing the pertinent status and the presence, if appropriate, of abnormal situations from the safety/systems failure point of view.

*E: System Administrator:* Evaluator of the data showed by the OCC. Will operate accordingly to the UPTS current status and its environment, whether triggering a specific security protocol against failure or solving the different spurious scenarios that may occur.

---

### Algorithms involved

---

Despite its apparent disparity, the following Computation Paradigms and the Algorithmic Techniques described below fall within the spectrum of the Natural Computation Paradigm. As long as the investigation is currently on an early-medium stage, the nexus with the system of some of these paradigms, as well as their application to the system, are still being under investigation.

As the accustomed reader will surely intuit, the algorithmic entities attached to this paradigm have, as its main base, the logic associated to phenomena present in nature, as well as the logic associated with the genetic-molecular base of the living beings, thus. As it can be read in [Rozenberg, 2012], we can



formally define Natural Computing as the set of computing techniques that circumscribe to, at least, one characteristic defined within the following group:

- \* Obtain its base from observing nature, establishing a computing simile.
- \* Base its reasoning in the use of the computers in order to synthesize natural phenomena.
- \* Use natural materials, from the logic or physical point of view, like DNA strings or chromosomes, in order to achieve its computational processes.

### Genetic Algorithms

As stated by Charles Darwin in his opus magnum *On the Origin of the Species* [Darwin, 1859], from immemorial times living beings have been forced to a continuous evolutive process looking for survival. Every single specie evolves from a common antecessor looking for the adaptation to its environment and survive, following the process named *natural selection*. In a parallel way, *Genetic Algorithms* (GA hereinafter) follows the same pattern, trying to evolve a *population*. Thus, as it can be extracted from John H. Holland’s *Adaptation in Natural and Artificial Systems* [Holland, 1975], a GA can be formally defined as a set of ordered instructions, that aims to achieve an specific problem, which are based on the genetic-molecular base of the evolutive process of the living beings. It is remarkable that, despite the paternity of the GA is attributed to Prof. Holland, his sublime work means the colophon to the investigator cycle started by the distinguished Gregor Mendel in 1865, with his laws stated in *Experiments in Plant Hybridization* [Mendel, 1865], based on the investigation over *Pisum Sativum*. In his publication, Mendel describes, using this specific pea variation, the basic rules related to the characteristics transmission between individuals through genetic inheritance. Actually, a GA has the objective of evolution certain specimens that set a *population*. In order to chase this goal, the GA uses random operations that establish a simile with the natural processes related to biological evolution. These methods, called *genetic operators*, are the following:

- Selection: In this operator, the GA chooses individual genomes from the population in order to start a later breeding process. Selection can be made by means of various techniques, as seen in *A Comparison of Selection Schemes Used in Evolutionary Algorithms* [Blickle, 1996]. These techniques can be Roulette-Wheel Selection, Selection by Truncation, Selection by Ranking or Selection by Tournament, to quote a few of them.
- Crossover: Process whereby a variation in the chromosomes is done from a generation towards the following one. It is remarkable that, following the natural simile, the crossover mocks the sexual reproduction of the living beings. Letting a binary string be the information to be represented, there are several crossover techniques, and they all produce permutations in the chromosome. Seeing the chromosome as a set of alleles, the technique of *crossover in a point* can be an illustrative example; as

shown in the following figure, once a bit within the chromosome is selected, every successive allele is exchanged between a chromosome and its pair, generating a new offspring in the process (Figure 3):



Figure 3. Genetic operator crossover in a point between one-byte alleles (Source: University of Amsterdam – Faculty of Sciences, Department of Computer Sciences)

- Mutation: Variation within the genotype of a living being. Represents the action of the mutagens present in the ecosystem. It is remarkable that the genetic unit able to mutate is the gene, atomic, inheritable unity of data that builds up an individual's DNA.
- Recombination: Process whereby a DNA portion is cleaved in order to provide its further union to a different genetic material molecule. It is important to note that this action provokes different genetic permutations in a specie regarding its predecessors, producing chimeric alleles. This advantage makes the sexual reproduction possible between living beings, while avoiding Muller's ratchet<sup>1</sup>.

### Ant Colony Optimization

As Marco Dorigo and Gianni Di Caro establishes several times along [Dorigo, 1992], Ant Colony Optimization (ACO hereinafter) is the name that refers to a multi-agent paradigm where every agent's behaviour is inspired on the ants idiosyncrasy when searching for livelihood. The algorithms that fall within this classification are based in Goss Experiment, using an *Iridomyrmex humilis* colony. In this experiment, the ant nest is connected to a livelihood source by means of two different paths, where one is longer than its counterpart, as the following Figure 4 shows:

<sup>1</sup> Named after its discoverer, Hermann Joseph Muller, is the process by which the different genomes of an asexual population accumulate deleterious mutations in an irreversible manner, that may result in the irrevocable extinction of the specie.

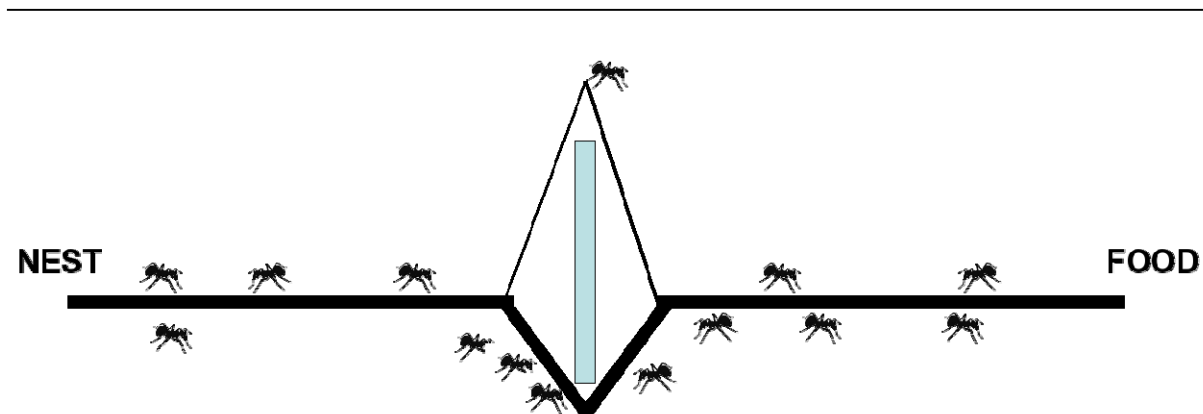


Figure 4. Goss Experiment representation (Source: BioDat Research Group - University of Karlovo, Prague)

After allowing the ants to freely move themselves along the scenario, it can be seen that, after an initial moments, they always choose the optimal, shorter path to the livelihood source. It is remarkable that, as well, this experiment demonstrates that a route selection probability is directly proportional to the length difference between both paths.

After studying the results thrown by Goss Experiment, a question arises; How do all the ants know what is the shortest path? The answer to this question is based on the concept known as stigmergy. The aforementioned concept alludes to those collaboration protocols, through the physical medium, where the different components collaborate due to the accumulation of objects or magnitudes in the environment, such pheromones or humidity. This concept is, precisely, the main tool within ant's communication; as the ants go back and forwards to the livelihood source, they deposit a chemical substance called pheromone. As it happens in several species, this substance provokes specific reactions and behaviour in the individual counterparts, allowing this, on this case, to know what the shortest path is.

It is remarkable that the directive that makes each ant  $k$ , placed in the  $i$ -th node, using a pheromone trail  $\tau_{ij}$  in order to calculate the probability it has to use to chose a node  $j$  that belongs to  $N$ , as well as the following node where it has to move along, where  $N_i$  constitutes the set of nodes adjacent to  $i$ , is given by the formula:

$$p_{ij}^k \begin{cases} \tau_{ij} & \text{if } j \in N_i \\ 0 & \text{if } j \notin N_i \end{cases} \quad (1)$$

### Particle Swarm Optimization

Since the dawn of science, many scientific have been intrigued by a movement, as elegant as optimal, present in nature: The harmonious synchrony in bird flocks and fish shoals, where the individuals are

able to move without even rub with each other, despite the hundreds, thousands of elements in certain cases, of individuals present in these sets. Thanks to scientific investigation, it has been demonstrated that, apart from this optimal movement, these animals present certain *swarm patterns* in their behaviour.

Concretely, it is important to highlight the hyperbolic interest of Grenander & Heppner on their opus magnum *A stochastic nonlinear model for coordinated bird flocks* [Grenander, 1990], where both zoologists synthesize their investigation referred to the nature-hidden directives that mark the asynchronous movement of the bird flocks, changing its direction suddenly in the presence of predators and tacitly regrouping, among other interesting abilities. In the same line, Reynolds *Flocks, herds and schools: a distributed behavioural model* [Reynolds, 1987] stands out, aiming to the study of the interesting choreography that birds deploy.

Clustering the aforementioned references as base, the Particle Swarm Optimization (PSO, hereinafter) paradigm is known as the technique that pretends to optimize a problem due to a meta-heuristic strategy, which is, due to the iterative *trial of improving* a candidate solution with regards to a pre-stipulated quality criteria. Thus, in a way that reminds to GA, PSO optimizes a problem starting from a set of candidate solutions, typically particles over the space, moving them along through the searching space without forgetting the premises of PSO mathematical base, which involves the position and the speed of the particles. As it can be inferred, the technique mimetizes the group behaviour of the aforementioned living beings, where each individual movement is influenced by the best *local* position known, while, in a parallel way, the *swarm* maintains a best *global* position known. This best global position is updated by the best position known by all the individuals in the swarm, fact that will guide the set to move searching for the best global position.

PSO adopts a tiny number of postulations along its execution process, exploring a mastodontic search space. Despite from that, PSO is a meta-heuristic, so it is not possible to adamantly ensure that the algorithm is going to find an optimal solution of the problem for every single case. In a more mathematical, accurate way, PSO does not use the *gradient* of the tackled problem, which means that this technique does not require the problem to be differentiable, as well as it happens in typical optimization methodologies such Cuasi-Newtonian methods or Gradient Descent. Thus, PSO can be used, enjoying a high success rate, in optimization problems that are especially non-regular, where there is certain ambient *noise*, or those presenting a dynamic, changing-over-time behaviour. PSO algorithm pseudocode can be stated as following, being *S* the swarm, *b\_low* and *b\_up* the preselected ranges and  $\omega$ ,  $\varphi$  parameters to be set:

```

for each (particle within S){
  position = generateRandomValue(S[i], b_low, b_up);
  position = bestKnownPositionByParticle(S[i]);
  if( f(p) < f(g) ){
    bestGlobalPosition = position;
  }
  speed = generateRandomSpeed(b_low - b_up, b_low - b_up)
}

while(!stopCriteria){
  for each (particle within S){
    for each (dimension within d){
      first_op =  $\omega v_{(i,d)} + \varphi_p * r_p$ ;
      second_op =  $(p_{(i,d)} - x_{(i,d)})$ ;
      third_op =  $\varphi_g * r_g * (g_d - x_{(i,d)})$ ;
      d[i] = first_op * second_op + third_op;
    }
    Position += d[i];
    if( f(xi) < f(pi) ){
      bestParticleLocalPosition = xi;
      if( f(pi) < f(g) ){
        bestGlobalPosition = pi;
      }
    }
  }
}
return bestGlobalPosition;
}

```

---

### Application to the smart cities realm and potential results

---

As surely the reader can imagine after studying this paper, the wide spectrum of applications that can be extracted from Natural Computing and applied to the smart city is, almost solely, bounded by wit limits. When it comes to a city's realm, the citizens can be seen as *particles*, thus they conform a *swarm*, atomic work element of the Natural Computing paradigm.

On the one hand, the efforts under the current investigation are being driven into the ACO Algorithm spectrum: Even efficient, ant pheromone is simple, primitive; it only marks the shortest path to the livelihood source, but; what if this pheromone concept is *extended* to a *super-pheromone*? A super-pheromone will store dissociated data of a person (i.e.: age, gender, education level, etc), thus, it will be possible to know which *person profile* is transiting for each UPTS section by seeing the user as an ant. More concretely, by applying the schema shown in [Figure 2. Overview of the currently under development system] under the epigraph [3. Topology of the system], the UPTS will acquire conscience, knowing who is circulating where, and consequently showing publicity screens according the relevant information for the public. (For instance, it will be more effective to show the publicity related to a new videogame near an institute area when the train is crowded by young people, while a new credit card

with certain bonuses will be more appropriate in the UPTS section beneath the financial area of the city.)

On the other hand, GA paradigm is being used as a way to *evolve* a route instead of a chromosome population: by means of a Smartphone application, users will be able to quickly know the best route between two points in the UPTS, as well as backup routes in case of systems breakdown. PSO can be used for studying the data retrieved in the Operations and Control Center (OCC). This will make possible to optimize the system by applying a statistical investigation over the data, detecting statistical outliers and acting in consequence. It is remarkable that the investigation regarding this slope and other Natural Computing paradigms is in an early stage, thus new applications are susceptible to emerge.

---

## Conclusions

In this paper, a newfangled scheme for endowing intelligence to a city UPTS is given, chasing the transition of the city to a *smart city*. In this approach, Natural Computing paradigm will be applied to the system, after a deep investigation that aims to improve the involved paradigms, if possible. Despite the investigation is still being in an early stage, the system is likely to improve the data gathering related to the UPTS in a mastodontic way, allowing the pertinent authorities to improve the system and even monetize the information gathered by the system under development. Moreover, users will be able to enjoy a better use of UPTS, knowing alternative routes in case of systems breakdown and being able to travel in an efficient way.

---

## Acknowledgements

We would like to thank Technical University of Madrid (Universidad Politécnica de Madrid – <https://www.upm.es>) for the granted permission in order to use the Center of Calculation and Communications for the simulations involved within the realm of this project.

---

## Bibliography

- [Blickle, 1996] Blickle, T, Thiele, L. (1996). A Comparison of Selection Schemes Used in Evolutionary Algorithms. *Evolutionary Computation* 4 (4): 361–394, 1996
- [CEOE, 2015] Confederación Española de Organizaciones Empresariales (CEOE). Acciones prioritarias para el desarrollo de las Smart Cities en España (Priority actions for the development of the Smart Cities in Spain). Smart Cities Committee, 4-8, Madrid, 2015.
- [Darwin, 1859] Darwin, C. *On the Origin of Species*, London, 1859.
- [Dorigo, 1992] Dorigo, M, Di Caro, G. *The Ant Colony Optimization Meta-Heuristic*, 1999.

- [Grenander, 1990] Grenander, L. Heppner, C. *A stochastic nonlinear model for coordinated bird flocks*, 1990.
- [Holland, 1975] Holland, J. *Adaptation in natural and artificial systems*. Cambridge, Mass.: MIT Press, 1975.
- [Hollands, 2008] Will the real smart city please stand up?. *City*, [online] 12(3), pp.303-320. Available at: [http://www.tandfonline.com/doi/abs/10.1080/13604810802479126#.Vw01M\\_mLSHs](http://www.tandfonline.com/doi/abs/10.1080/13604810802479126#.Vw01M_mLSHs) [Accessed 2 Apr. 2016].
- [Mendel, 1865] Mendel, G. *Experiments in plant hybridisation*. Cambridge: Harvard University Press, 1865.
- [MTA, 2012] Metropolitan Transportation Authority (MTA). Annual Subway Ridership [online]. Available at: <http://web.mta.info/nyct/facts/ffsubway.htm>[Accessed 5 Apr. 2016].
- [Reynolds, 1987] Reynolds, H. *Flocks, herds and schools: a distributed behavioral model*, 1987.
- [Rozenberg, 2012] Rozenberg, G., Bäck, T. and Kok, J. *Handbook of natural computing*. Berlin: Springer, 2012.

---

#### Authors' Information

---



**Clemencio Morales** – *Computer Engineer, Master on Web Engineering, PhD Candidate at Technical University of Madrid (Spain), Madrid, Spain; e-mail: [mail@clemenciomorales.com](mailto:mail@clemenciomorales.com)*

*Major Fields of Scientific Research: Natural Computing, Ant Colony Optimization, Smart Cities, Software Engineering.*



**Luis Fernando de Mingo** – *Professor and investigator at Technical University of Madrid (Spain), Madrid, Spain; e-mail: [fernando.demingo@upm.es](mailto:fernando.demingo@upm.es)*

*Major Fields of Scientific Research: Natural Computing, Software Engineering, Ubiquitous Programming.*

## Clustering using Particle Swarm Optimization

Nuria Gómez Blas, Octavio López Tolic

**Abstract:** Data clustering has been a well-studied research field for a long time. One of the latest trends in this area is the application of Particle Swarm Optimization (PSO) in clustering which has good potential for improvements. This paper presents an approach to using Particle Swarm Optimization to cluster data. It is shown how PSO can be used to find the centroids of a user specified number of clusters. Results show that PSO clustering techniques have much potential.

**Keywords:** Social intelligence, Particle swarm optimization, Swarm computing, Clustering algorithms.

**ACM Classification Keywords:** 10010147.10010178 Computing methodologies Artificial intelligence, 10010147.10010257 Computing methodologies Machine learning, 10010147.10010257.10010293.10010294 Computing methodologies Neural networks.

**Conference topic:** Information Modelling, Information Systems, Applied Program Systems.

**MSC:** 68Q32 Computational learning theory, 68T05 Learning and adaptive systems.

---

### Introduction

---

Cluster analysis was originated in anthropology by Driver and Kroeber in 1932 and introduced to psychology by Tryon [1939] and famously used by Cattell [1943] for trait theory classification in personality psychology. Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy, botryology and typological analysis. The subtle differences are often in the usage of the results: while in data mining, the resulting groups are the matter of interest, in automatic classification the resulting discriminative power is of interest. This often leads to misunderstandings between researchers coming from the fields of data mining and machine learning, since they use the same terms and often the same algorithms, but have different goals.

The notion of a cluster cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms Estivill-Castro [2002]. There is a common denominator: a group of data objects. However, different researchers employ different cluster models, and for each of these cluster models again different algorithms can be given. The notion of a cluster, as found by different algorithms, varies significantly in its properties. Understanding these cluster models is key to understanding the differences between the various algorithms. Typical cluster models include:

- Connectivity models: for example, hierarchical clustering builds models based on distance connectivity.
- Centroid models: for example, the k-means algorithm represents each cluster by a single mean vector.
- Distribution models: clusters are modeled using statistical distributions, such as multivariate normal distributions used by the Expectation-maximization algorithm.
- Density models: for example, DBSCAN and OPTICS defines clusters as connected dense regions in the data space.
- Subspace models: in Bicustering (also known as Co-clustering or two-mode-clustering), clusters are modeled with both cluster members and relevant attributes.



- Group models: some algorithms do not provide a refined model for their results and just provide the grouping information.
- Graph-based models: a clique, that is, a subset of nodes in a graph such that every two nodes in the subset are connected by an edge can be considered as a prototypical form of cluster. Relaxations of the complete connectivity requirement (a fraction of the edges can be missing) are known as quasi-cliques, as in the HCS clustering algorithm.

Data clustering is a popular approach of automatically finding classes, concepts, or groups of patterns. It seeks to partition an unstructured set of objects into clusters (groups). This implies wanting the objects to be as similar to objects in the same cluster and as dissimilar to objects from other clusters as possible. Clustering has been applied in many areas including biology, medicine, anthropology, marketing and economics. Clustering applications include plant and animal classification, disease classification, image processing, pattern recognition and document retrieval. Clustering techniques have been applied to a wide variety of research problems.

A clustering is essentially a set of such clusters, usually containing all objects in the data set. Additionally, it may specify the relationship of the clusters to each other, for example, a hierarchy of clusters embedded in each other. Clusterings can be roughly distinguished as: hard clustering – each object belongs to a cluster or not and soft clustering (also: fuzzy clustering) – each object belongs to each cluster to a certain degree (for example, a likelihood of belonging to the cluster).

Clustering algorithms can be grouped into two main classes of algorithms, namely supervised and unsupervised. With supervised clustering, the learning algorithm has an external teacher that indicates the target class to which a data vector should belong. For unsupervised clustering, a teacher does not exist, and data vectors are grouped based on distance from one another. Many unsupervised clustering algorithms have been developed. Most of these algorithms group data into clusters independent of the topology of input space. These algorithms include, among others, K-means [Kanungo et al. \[2002\]](#), ISODATA [Memarsadeghi et al. \[2007\]](#), and learning vector quantizers (LVQ) [Fausett \[1994\]](#).

This paper presents the fundamentals of Particle Swarm Optimization algorithm and also it introduces the application to clustering problems.

Particle swarm optimization (PSO) [Kennedy and Eberhart \[1995\]](#) is a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. It solves a problem by having a population of candidate solutions, here dubbed particles, and moving these particles around in the search-space according to simple mathematical formulae over the particle's position and velocity. Each particle's movement is influenced by its local best known position but, is also guided toward the best known positions in the search-space, which are updated as better positions are found by other particles. This is expected to move the swarm toward the best solutions.

---

### Particle Swarm Optimization Model

---

Particle swarm optimization (PSO) is a heuristic optimization technique, originally developed by James Kennedy and Russell C. Eberhart in 1995. PSO is considered to fall within the branch of evolutionary computing, which is commonly referred as swarm intelligence. Evolutionary Computation is a collection of methods that simulate evolution by using computers, including genetic programming, evolution strategies, Swarm Optimization, Artificial Life and other related fields. Evolutionary computation and genetic algorithms share the technique for generating populations in a stochastic way. Furthermore, this technique is used to generate new generations too.

The algorithm used by PSO is inspired by the collective work of swarms, specifically in the social organization of bird flocks and fish schools. These social organizations take on a collective behavior that comes from communication and cooperation among its members. In the algorithm, the individuals of the population are treated as particles or searching points, which have a position and velocity (the velocity vector indicates where it is headed). It move into an area of dimension  $n$ . The particle that gets better assessment is the *leader*, in the sense that the whole swarm

will follow it. However any individual could change their orientation. In that case the leader passes leadership to the one that changed the orientation. The *leader* can either be global leader in the whole swarm or local in a part of it.

The emulation of these organized societies has been successful in discrete and continuous optimization problems. Initially, Kennedy and Eberhart developed a standard algorithm, which has undergone several improvements, it is a fairly simple algorithm and widely used today.

The particles move in the search space by using a combination of the best individual solution found by the particle and the best found by any of the neighboring particles. For every iteration we evaluate the performance of each particle.

The three main operations Kennedy and Eberhart [1995] algorithms use are: To assess, to compare and to imitate. Evaluation is one of the most important features of some living organisms; they evaluate learn and evolve. Besides, organisms analyze their neighbors and imitate only those who have better performance. In general, these three operations can be applied to simple social beings and to computer programs. This enable them for solving complex problems.

Each particle  $i$  is related to three vectors: position:  $x^{(i)} = (x_1, x_2 \dots, x_d)$ , the best position of its history:  $p^{(i)} = (p_1, p_2 \dots, p_d)$  and speed  $v^{(i)} = (v_1, v_2 \dots, v_d)$ , given a space of dimension  $d$ . Initially values related to particles are generated randomly, then these particles move through a search space by using a system of equations that is updated for each iteration. The intention is to find the best solution. Each particle moves to the more successful neighbor, influencing the other particles. The algorithm updates the swarm at each step. It also changes the position and velocity of each particle and it applies the following rules:

$$v_d^{(i)} = v_d^{(i)} + c_1 \epsilon_1 (p_d^{(i)} - x_d^{(i)}) + c_2 \epsilon_2 (g_d^{(i)} - x_d^{(i)}) \quad (1)$$

$$x_d^{(i)} = x_d^{(i)} + v_d^{(i)} \quad (2)$$

$v_d^{(i)}$  is the  $d$ -th component of the velocity of particle  $i$ ,  $x_d^{(i)}$ , the  $d$ -th component of vector position of the particle  $i$ ,  $c_j$ , ( $j = 1, 2$ ) is a constant value (obtained from experimental results) and  $\epsilon_1$  and  $\epsilon_2$  are independent random numbers uniformly distributed in  $[0,1]$ . They are generated for every update,  $p_d$  is the  $d$ -th component of the improved performance of the particle in its history, and  $g_d$  is the  $d$ -th component of the particle with the best position found between neighboring particles, throughout its history. This neighborhood is defined according to the topology of the particle system. The factor  $c_1$  is known as the factor of personal or cognitive learning and the factor  $c_2$  as the social learning factor. Both factors have much influence on the rate of convergence of the optimization process.

Mathematical models developed for PSO vary according to how the particles interact with their neighbors, this is known as the topology of the system, which can be understood as the way the swarm of particles organizes itself. Early models of PSO used a Euclidean neighborhood. However the number of operations were high; in order to reduce the number of operations, a new mathematical model was developed. In this mathematical model neighborhoods were not related to the location of the particle; these are called local neighborhoods or *lbest models*. They can be also global or and *gbest models*. Originally *gbest* had better performance, but recent research has also shown good results in some problems when using the model *lbest* by adding some improvements to the algorithm.

In PSO, neighborhood is understood as the set of particles related to a given particle. This relationship influences the search capability and convergence. In the ring-type topology, each particle communicates only with  $n$  neighbors,  $n/2$  on each side. Ring topology presented is the simplest:  $n = 2$ . This means that only two neighboring particles are evaluated. In the wheel topology all information is concentrated in a central particle. In the star type each particle is related to all the particles of the swarm.

The standard Particle Swarm Optimization schema is described as follows, see algorithm 1.

It begins with a population of particles with position and velocity randomly assigned in the search space. With regard to calculating the velocity of the particles, we consider a maximum value as restriction. This is called  $V_{max}$ . This is done because an explosion might happen, since the velocities could be increased quickly. The selection of the

**Algorithm 1** Standard Particle Swarm Optimization Algorithm

---

```

1: Create initial population of particles  $x^i$ 
2: loop
3:   for each particle of the swarm do
4:     if  $f(x) > f(p)$  then
5:       for  $d = 1 \rightarrow D$  do
6:          $p_d = x_d$ 
7:       end for
8:     end if
9:      $g = i$ 
10:    for  $j \in J$  do
11:      if  $f(p_j) > f(p_g)$  then
12:         $g = j$ 
13:      end if
14:    end for
15:    for  $d = 1 \rightarrow D$  do
16:       $v_d(t) = v_d(t - 1) + c_1\epsilon_1(p_d - x_d(t - 1)) + c_2\epsilon_2(g_d - x_d(t - 1))$  1
17:       $x_d = x_d + v_d$ 
18:    end for
19:  end for
20: end loop

```

---

values  $V_{max}$  is difficult to determine. They will be chosen by trial and error. In very large spaces, large values are usually selected in order to ensure proper exploration. This is justified since a large inertia weight facilitates exploration in new areas in the global search space, while a small one facilitate the exploration in a local area. In the case of small spaces, small values are required to prevent the explosion. As for the size of the particle population, empirical results have shown that good results are not always obtained by increasing the number of particles of the population. Some examples have been influenced by that but others have not.

The particle swarm is actually more than just a collection of particles. A particle by itself has almost does not solve any problem; progress takes place only when they i.e. the particles interact. Populations are organized according to some sort of communication structure or topology. This is often thought of as a social network. The topology typically consists of bidirectional edges connecting pairs of particles. It is like the alphabet  $j$  appearing in  $i$ 's neighborhood, and likewise  $i$  in  $j$ 's neighbour. Each particle communicates with other particles and is affected by the best point found by any member of its topological neighborhood.

**Model variants**

Regarding the PSO algorithm, different variants have been developed, aimed at speeding up the convergence of it. In addition to the unconstrained optimization problem in discrete or continuous variable, the multi target problem and the constrained problem have been addressed. We have also developed hybrid optimization techniques, PSO technique has been tested with good results for training Artificial Neural Networks. When applying the method of Back Propagation, we are able to find appropriate weights that minimize an error function through a succession of iterations. On the other hand, by applying the PSO technique, the weights found are more efficient just by making small modifications to the algorithm. The new guidelines are aimed at avoiding PSO stagnation of the local optimal solutions.

Kennedy et al. [2001] proposed adjustments to the velocities of the particles by using a factor  $w$  called *inertial weight*. This factor utilizes the inertia of the particles in the process of friction when they are moving. This modification in the algorithm is done to control the search space. In order to do that it must change (3). The large inertia weight

makes the global search easier; however small inertia weight does not improve local search. That is why the initial value is greater than 1.0 to promote global exploration, and then gradually decreases to obtain more refined solutions. The algorithm decreases linearly at each iteration. Moreover, the use of inertial weight removes the restriction  $V_{max}$  on the velocity.

$$v_d^{(i)} = wv_d^{(i)} + c_1\epsilon_1(p_d^{(i)} - x_d^{(i)}) + c_2\epsilon_2(g_d^{(i)} - x_d^{(i)}) \quad (3)$$

In each iteration, inertia weight decrease linearly through the following expression:

$$w = w_{max} - (w_{max} - w_{min})\frac{g}{G} \quad (4)$$

$g$  is the index of the generation,  $G$  is the maximum number of iterations previously determined,  $w_{max}$  is a value greater than 1, and  $w_{min}$  a value under 0.5. This variation of the method has proven to accelerate convergence.

Clerc and Kennedy [2002] obtain another variation in the speed calculation. A constriction factor  $\chi$  is introduced with that purpose, This factor depends on the constants that are used when calculating speed. This factor affects the formula (1) The aim is to avoid the explosion of velocity:

$$v_d^{(i)} = \chi[v_d^{(i)} + c_1\epsilon_1(p_d^{(i)} - x_d^{(i)}) + c_2\epsilon_2(g_d^{(i)} - x_d^{(i)})] \quad (5)$$

$\chi$  is:

$$\chi = \frac{2}{|2 - \varphi - \sqrt{\varphi^2 - 4\varphi}|}, \quad \varphi = c_1 + c_2 = 4.1$$

The results are:  $\chi = 0.729$  and  $c_1=c_2=2.05$ . These parameters were obtained from performing several tests.

$\chi$  factor is similar to the inertial weight. This means that controlling the velocity with  $V_{max}$  is not required when  $\chi$  is used. Bratton and Kennedy [2007], analyzed the stability of this algorithm by using these values and by following a comparative study of both PSO algorithms (inertial weight and  $\chi$  factor). Both of them are mathematically equivalent, in particular the algorithm with constriction factor is a special case of the inertial weight. Moreover, Parsopoulos et al. [2001], combined both for problems with constraints and they obtained equally good results in several tests.

We observe that the convergence always becomes slower when problem size increase, so when it comes to high-dimensional problems, a larger number of iterations occurs. Researchers Uchitane and Hatanaka [2015]; Hatanaka et al. [2015] developed a PSO model, where velocity values are updated, by considering the rotation of the coordinate system. This model is aimed at problems of high dimensionality and it showed good results when applied to all functions of De Jong, (larger dimensions).

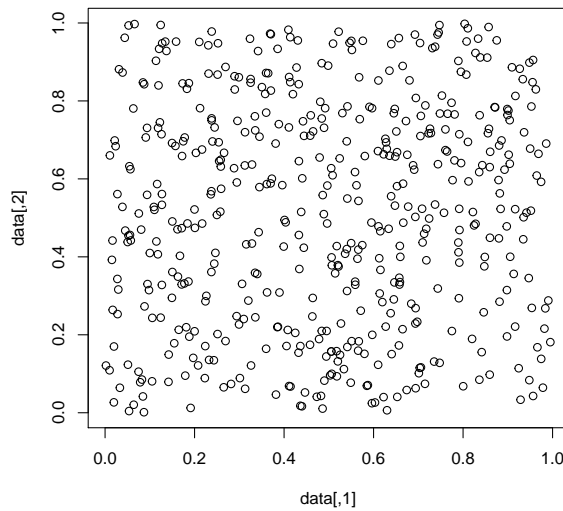
---

## Clustering using Particle Swarm Optimization

---

This section shows a first approach to data clustering problems using Particle Swarm Optimization. Pictures have been generated using R, Sweave and L<sup>A</sup>T<sub>E</sub>X. Algorithms implemented in this paper can be shown in listings 1 and 2. Data sets have been generated using an uniform distribution in square [0,1] with `runif` R function. Figure 1 shows used data in the clustering problem represented by figure 2.

In order to start with the PSO algorithm each particle is considered to represent the centroid of a given cluster. In the context of clustering, a single particle represents the cluster centroid vectors,  $p = (x, y)$  in the 2 dimensional space. When dealing with multiple clusters, let's say  $n$ , each particle is represented as a collection of clusters, that is  $p = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ . Using this approach particle dimension will increase when working according the number of clusters.

Figure 1: Uniform data in  $[0, 1]$  used in the PSO clustering problem.

The fitness of particles is easily measured as the quantization error. The fitness function of the data clustering problem is given as follows:

$$f = \frac{1}{1 + d} \quad (6)$$

The function  $f$  should be minimized. Where

$$d \text{ is the mean distance of data with respect to clusters} \quad (7)$$

One particle in the swarm represents one possible solution for clustering. Therefore, a swarm represents a number of candidate clustering solutions for the data set. At the initial stage, each particle randomly chooses  $k$  different data set from the collection as the initial cluster centroid vectors.

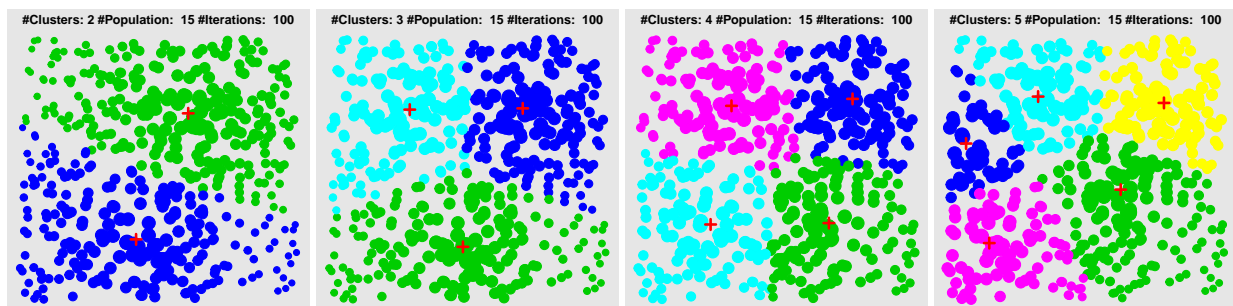
Figure 2: Clustering uniform data in  $[0, 1]$  square using PSO with different clusters.

Figure 2 shows obtained results with the PSO algorithm using different clusters, that is 2, 3, 4, and 5 clusters. We can observe that all centroids are well distributed.

Figures 3 and 4 shows other different distribution of initial data using different clusters.

---

## Final Remarks

One shortcoming of the PSO algorithm is the formation of number of small clusters which was overcome by introducing a one setp K-means operator that forces the small clusters into the bigger ones and finishes the

Figure 3: Clustering uniform data using PSO with different clusters.

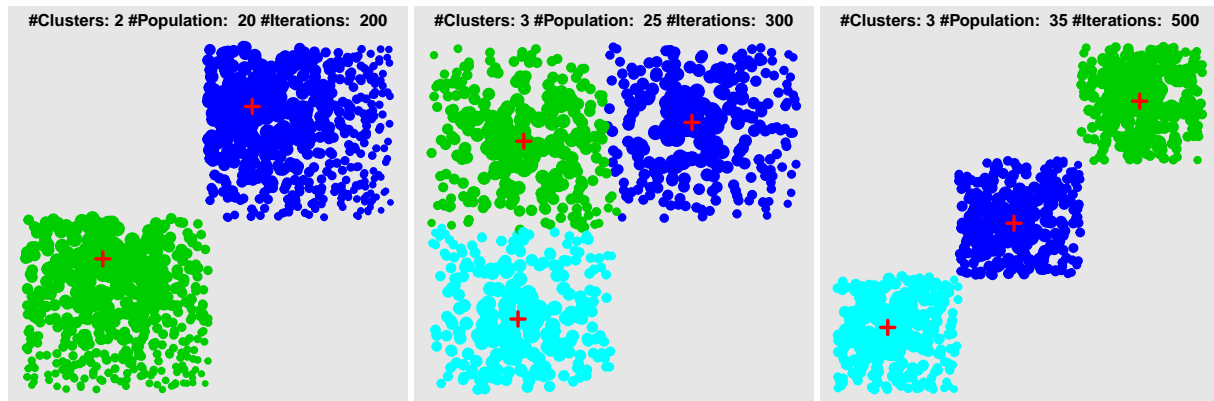
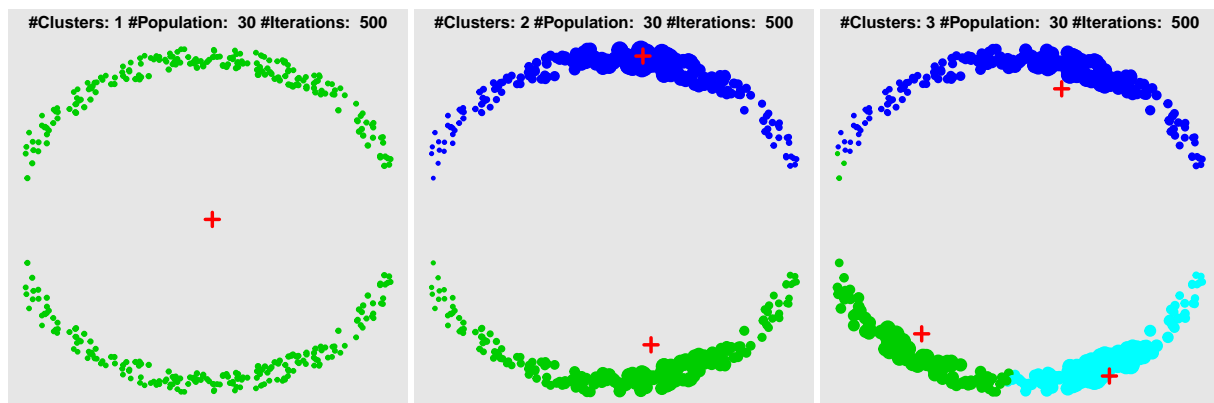


Figure 4: Clustering circle data using PSO with different clusters.



clustering operation [van der Merwe and Engelbrecht \[2003\]](#). But this goes against the spirit of the algorithm and the user will have to specify the minimum number below which the cluster will be considered as a small cluster. Another approach could be to introduce a second phase of PSO in which the parameters are reinitialized and positions reallocated. New force functions could be defined which will increase the attractive force between larger and smaller clusters and reduces the force between larger clusters. Particle swarm optimization provides number of future applications and improvements in the domain of clustering.

## Appendix: Listings of PSO clustering in R

Listings 1 and 2 show the implementation algorithms used in this paper. First one –listing 1– correspond to a general PSO algorithm while the second one –listing 2– corresponds to the clustering approach based on a PSO algorithm.

Listing 1: Basic Particle Swarm Algorithm. File `pso.R`.

```

fitness <- function(pop, calculate_fitness) {
  pop[,dimensions+1] <- calculate_fitness(pop[,1:dimensions]);
  return(pop)
}

pso <- function(verbose, individuals, dimensions, iterations, c1, c2, inertia, calculate_fitness) {
  if (verbose)
    cat ("Startint PSO\n");

  #####
  # Init population (actual + fitness + local_best + fitness)
  population <- array (runif(2*(dimensions+1)*individuals), c(individuals, 2*(dimensions+1)));

  # Calculate fitness
  population <- fitness (population, calculate_fitness);
  # Init local_best
  population[, (dimensions+2):(2*(dimensions+1))] <- population[,1:(dimensions+1)];
  # Init global_best

```

---



---

```

global_best <- population[max.col(t(population[,dimensions+1])),1:(dimensions+1)];

# Init velocity
velocity <- array (rnorm(individuals*dimensions), c(individuals, dimensions));
inertia_coeff <- inertia/iterations;

# Init historic data
fitness_historic <- array(0, iterations);
global_best_historic <- array(0, c(iterations,dimensions + 1));

for (iteration in 1:iterations) {
  uniform_coeffs <- runif(individuals*2);
  # Update velocity
  velocity <- inertia * velocity +
    c1 * uniform_coeffs[1:individuals] * (population[, (dimensions+2):(2*(dimensions+1)-1)] - population[,1:dimensions]) +
    c2 * uniform_coeffs[(individuals+1):(2*individuals)] * (global_best[1:dimensions] - population[,1:dimensions]);
  # Update inertia
  inertia <- inertia - inertia_coeff;

  # Update population with fitness
  population[,1:dimensions] <- population[,1:dimensions] + velocity;
  population <- fitness (population, calculate_fitness);

  # Update local_best
  for (individual in 1:individuals) {
    if (population[individual, dimensions+1] > population[individual, 2*(dimensions+1)]) {
      population[individual, (dimensions+2):(2*(dimensions+1))] <- population[individual, 1:(dimensions+1)];
    }
  }

  # Update global_best
  if (max (population[,dimensions+1] > global_best[dimensions+1])) {
    global_best <- population[max.col(t(population[,dimensions+1])),1:(dimensions+1)];
  }

  if (verbose)
    cat("Iteration ", iteration, " Fitness ", global_best[dimensions+1], "\n");

  # Keep historic fitness data
  fitness_historic[iteration] <- global_best[dimensions+1];

  # Keep historic global_best
  global_best_historic[iteration,] <- global_best;
}

#####
if (verbose)
  cat ("Finish PSO\n");

return(list(best=global_best, fitness_evol=fitness_historic, global_best=global_best_historic));
}

```

---

Please note that in listing 2 parameters iterations, individuals, clusters and data must be set to the desired configuration to solve a given problem. In our case some sample code to call these functions is the following, Sweave package and R have been used to generate a source  $\LaTeX$ :

```

data <- array(runif(1000), c(500,2),dimnames=list(NULL,paste('Dimension',1:2)));
individuals <- 15;
iterations <- 100;
clusters <- 2;
source('nclustering.R');

```

Listing 2: Clustering Algorithm using PSO. File nclustering.R

---

```

#### N-Clustering sample
source('pso.R')

#### global parameters
# iterations <- 1;
# individuals <- 20;
# clusters <- 1;
# data <- array(runif(1000)*2 -1 , c(500,2));

#### general parameters
clusters_dimension <- 2;
dimensions <- clusters_dimension*clusters;
c1 <- 2;
c2 <- 1;
inertia <- .8;

# fitness function in clustering process
clustering_fitness <- function(ind) {
  val <- array(0, c(length(ind[,1]),1));
  for(individual in 1:length(ind[,1])) {
    dist <- 0;

```



---



---

```

for(i in 1:length(data[,1])) {
  dist_vector <- (data[i,]-ind[individual,])^2;
  dist_min <- array(0, clusters);
  for (j in 1:clusters) {
    dist_min[j] <- sqrt(sum(dist_vector[((j-1)*clusters_dimension + 1):(j*clusters_dimension)]));
  }
  dist <- dist + min(dist_min)/length(data[,1]);
  val[individual] <- 1 / (1 + dist);
}
return(val)
}

res <- pso(FALSE, individuals, dimensions, iterations, c1, c2, inertia, clustering_fitness);

# setup margins and title size
par(mar=c(0.2,0.2,2,0.2),bg=rgb(0.9,0.9,0.9), cex.main=1.5);

# plot data
plot(data[,1], data[,2], cex= 0.1, xlab='', ylab='',lwd =1, axes=FALSE);
title(paste("#Clusters:", clusters, "#Population:", individuals, "#Iterations: ", iterations));

distribution <- array(0,c(1,clusters), dimnames = list("Number of elements:",paste("Cluster",1:clusters)));

for(i in 1:length(data[,1])) {
  dist_vector <- (data[i,] - res$best[1:dimensions])^2;
  dist_min <- array(0, clusters);
  for (j in 1:clusters) {
    dist_min[j] <- sqrt(sum(dist_vector[((j-1)*clusters_dimension + 1):(j*clusters_dimension)]));
  }
  dist_min <- 1 / (1 + dist_min);
  winner <- max.col(t(dist_min));
  distribution[winner] <- distribution[winner] + 1;
  points(data[i,1], data[i,2],lwd=2,cex=(4*max(dist_min)^2), pch=19, col = winner+2);
}

for (i in 1:clusters)
  points(res$best[((i-1)*clusters_dimension)+1], res$best[(i)*clusters_dimension], col='red', pch = 3,lwd=5, cex=2);

```

---

## Bibliography

---

- Bratton, D. and Kennedy, J. (2007). Defining a standard for particle swarm optimization. In *2007 IEEE Swarm Intelligence Symposium, SIS 2007, Honolulu, Hawaii, USA, April 1-5, 2007*, pages 120–127.
- Cattell, R. (1943). *The Description of Personality: Basic Traits Resolved Into Clusters*. American psychological association.
- Clerc, M. and Kennedy, J. (2002). The particle swarm - explosion, stability, and convergence in a multidimensional complex space. *IEEE Trans. Evolutionary Computation*, 6(1):58–73.
- Estivill-Castro, V. (2002). Why so many clustering algorithms: A position paper. *SIGKDD Explor. Newsl.*, 4(1):65–75.
- Fausett, L., editor (1994). *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.
- Hatanaka, T., Chopra, N., and Fujita, M. (2015). Passivity-based bilateral human-swarm-interactions for cooperative robotic networks and human passivity analysis. In *54th IEEE Conference on Decision and Control, CDC 2015, Osaka, Japan, December 15-18, 2015*, pages 1033–1039.
- Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., Wu, A. Y., Member, S., and Member, S. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:881–892.
- Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of IEEE International Conference on Neural Networks*, volume IV, pages 1942–1948.
- Kennedy, J., Eberhart, R., and Shi, Y. (2001). *Swarm Intelligence*. Morgan Kaufman.
- Memarsadeghi, N., Mount, D. M., Netanyahu, N. S., and Moigne, J. L. (2007). A fast implementation of the isodata clustering algorithm. *Int. J. Comput. Geometry Appl.*, 17(1):71–103.



- 
- 
- Parsopoulos, K., Plagianakos, V. P., and Magoulas, G. D. (2001). Stretching technique for obtaining global minimizers through particle swarm optimization. In *Proceedings of Particle Swarm Optimization Workshop*, pages 22–29.
- Tryon, R. (1939). *Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality*. Edwards brother, Incorporated, lithoprinters and publishers.
- Uchitane, T. and Hatanaka, T. (2015). A study on multi-objective particle swarm model by personal archives with regular graph. In *IEEE Congress on Evolutionary Computation, CEC 2015, Sendai, Japan, May 25-28, 2015*, pages 2685–2690.
- van der Merwe, D. W. and Engelbrecht, A. P. (2003). Data clustering using particle swarm optimization. In *Evolutionary Computation, 2003. CEC '03. The 2003 Congress on*, volume 1, pages 215–220 Vol.1.

---

### Authors' Information

---



**Nuria Gómez Blas** - Dept. de Sistemas Informáticos, Escuela Técnica Superior de Sistemas Informáticos, Universidad Politécnica de Madrid, Crta. de Valencia km. 7, 28031 Madrid, Spain; e-mail: [nuria.gomez.blas@upm.es](mailto:nuria.gomez.blas@upm.es)  
Major Fields of Scientific Research: Bio-inspired Algorithms, Natural Computing



**Octavio López Tolic** - Dept. de Sistemas Informáticos, Escuela Técnica Superior de Sistemas Informáticos, Universidad Politécnica de Madrid, Crta. de Valencia km. 7, 28031 Madrid, Spain; e-mail: [oa@etsisi.upm.es](mailto:oa@etsisi.upm.es)  
Major Fields of Scientific Research: Bio-inspired Algorithms, Natural Computing

## On a public key encryption algorithm based on Permutation Polynomials and performance analyses

Gurgen Khachatryan Martun Karapetyan

**Abstract:** In this paper a modification of public key encryption system presented in [Khachatryan, Kyureghyan, 2015] and performance analysis are presented. As described in [Khachatryan, Kyureghyan, 2015], the permutation polynomial  $P(x)$  is declared to be a public polynomial for encryption. A public key encryption of given  $m(x)$  is the evaluation of polynomial  $P(x)$  at point  $m(x)$  where the result of evaluation is calculated via so called White box reduction, which does not reveal the underlying secret polynomial  $g(x)$ . Our analysis have shown that an attacker may acquire some information about the message, having its cipher-text, in case of using certain values of  $P(x)$ . So either those values of  $P(x)$  must be avoided, or the modification presented in this paper must be used. Our implementation's performance was compared to RSA-2048 implementation of CryptoPP library and it was 3.75x and 133x faster on encryption and decryption operations respectively.

**Keywords:** Permutation polynomials, Public-key encryption, White box reduction..

**ACM Classification Keywords:** E.3 DATA ENCRYPTION - Public key cryptosystems

**Conference topic:** Cryptographic methods and protocols, Reliable and Secure Telecommunications;

**MSC:** 11T71, 94A60

---

### Introduction

---

Let  $GF(q)$  be the finite field with  $q$  elements, where  $q$  is a prime or power of a prime. A polynomial  $f(x)$  over  $GF(q)$  is called a permutation polynomial if an equation  $f(x) = r$  for any  $r \in GF(q)$  has only one root in  $GF(q)$ . In [Khachatryan, Kyureghyan, 2015] a new class of permutation polynomials was presented and a public key system with white box implementation was provided. Its security relies on the problem of solving a polynomial equation over  $GF(2^n)$  when the field representation polynomial is unknown.

A pioneering work describing DH key exchange by Diffie and Hellman [Diffie, Hellman, 1976] was presented in 1976, which is based on the discrete logarithm problem (DLP). In 1978 another fundamental work by Rivest, Shamir and Adleman [Rivest, Shamir, Adleman, 1978], called RSA cryptosystem was presented, which is based on integer factorization problem. Another important development for public key cryptosystems was the invention of Elliptic curve cryptosystems [Miller, 1986] which are based on the algebraic structure of elliptic curves over finite fields.

In this paper we present a modification of the cryptosystem described in [Khachatryan, Kyureghyan, 2015] and performance analysis.

The paper is organized as follows: In section 2 the public key algorithm and white box implementation described in [Khachatryan, Kyureghyan, 2015] are presented in short. In section 3 modification of the public key system is presented. In section 4 implementation aspects and performance optimizations of the proposed system are discussed. Section 5 concludes the paper.

---

### Public Key encryption algorithm based on permutation polynomials

---

Let  $GF(q)$  be a finite field of characteristic  $p$ . For every permutation polynomial  $f(x)$  over  $GF(q)$ , there exists a unique polynomial,  $f^{-1}(x)$  over  $GF(q)$  such that  $f(f^{-1}(x)) = (f^{-1}(f(x))) = x$  called the compositional inverse of  $f(x)$ .

Let  $F(x) = \sum_{u=0}^n a_u x^u \in GF(2)$  be a primitive polynomial and  $P(x) = \sum_{u=0}^n a_u x^{2^u}$  be its linearized 2-associate. Elements of the  $GF(2^n)$  will be represented through a primitive polynomial  $g(x)$  over  $GF(2)$ . Then  $P(x)$  is a permutation polynomial and an algorithm for finding its compositional inverse  $P^{-1}$  was presented in [Khachatryan, Kyureghyan, 2015].

A primitive polynomial  $g(x)$  of degree  $n$  over  $GF(2)$  is used as the public key encryption private (secret) parameter and a permutation polynomial  $P(x)$  as the public parameter.

- a) Public key encryption: any message  $m(x)$  of the length  $n$  as an input (plaintext) and evaluation of the polynomial  $P(m(x)) = c(x) \text{ mod } g(x)$  as an output (ciphertext). The evaluation operation will be implemented via "White box" evaluation without revealing the polynomial  $g(x)$ . The output of public key encryption will be  $c'(x)$  based on "White box" tables explained later.
- b) Private key decryption: Given a ciphertext  $c'(x)$  calculate  $c(x)$ . Compute  $P^{-1}(c(x)) = P^{-1}(P(m(x))) = m(x) \text{ mod } g(x)$ . [Khachatryan, Kyureghyan, 2015]

The white box evaluation is used to calculate the value of  $c'(x)$  without revealing the value of modulo reduction polynomial  $g(x)$  in such a way, that the "owner" of the system can calculate  $c(x)$  from it. The evaluation procedure will be as follows: all possible residues  $x^N \equiv R_N(x) \text{ mod } g(x)$  for  $N = 2^i r$ , where  $i = 1, \dots, 128, r = 2k + 1, k = 0, 1, \dots, 63$  are biased by using random 64 secret polynomials  $L_0, L_1, \dots, L_{63}$  based on another secret polynomial  $L(X)$  which are only known to the "owner" of the system. All biased values for residues modulo polynomial  $g(x)$  are provided to the public in the following manner:

$$B_N(x) = ((R_N(x) \times L_0(x)) \text{ mod } L(x)) \oplus L_{k+1}(x) \tag{0.1}$$

for any  $N = 2^i(2k + 1)$ . Based on above explanation an encoding procedure will be as follows: The user calculates an evaluation result of the polynomial  $P(m(x))$  without any reduction, takes the polynomial  $R(x)$  that contains all terms of evaluation for the degrees not exceeding 127, and calculates the modulo two sum of nonzero terms  $B_N(x)$  denoted by  $\sum B_N(x)$  corresponding to nonzero terms of evaluation result exceeding  $N = 127$ .

An encrypted message  $c'(x)$  then contains two 16 byte vectors including  $R(x), \sum B_N(x)$  and another 8 byte vector  $B = (b_0, b_1, \dots, b_{63})$ , where  $b_k = 0$  if the number of nonzero terms with the same value  $k$  in evaluation result is even and  $b_k = 1$  otherwise for  $N = 2^i(2k + 1), k = 0, \dots, 63$ .

Decoding procedure by the "owner" of the system will be as follows: based on the value  $B = (b_0, b_2, \dots, b_{63})$  and vector  $\sum B_N(x)$  the "owner" calculates:

$$R(x) \oplus \left( \sum B_N(x) \oplus \sum_{i=1}^{128} b_i \times L_i(x) \right) \times (L_0(x))^{-1} = c(x) \text{ mod } L(x). \tag{0.2}$$

---

### Modification of the Public Key encryption algorithm

---

A public polynomial  $P(x) = \sum_{u=0}^n a_u x^{2^u}$  is used in the encryption algorithm described in [Khachatryan, Kyureghyan, 2015]. Our analyses showed, that if  $a_u = 0$  for all but one value  $v$  in  $u = 2..7$  then the attacker may gain some information about the plain-text message  $m(x)$  having the value of  $R(x)$ . For example is  $P(x) = x^{2^{127}} + x^2 + x$  is used, then if  $m(x) = \sum_{i=0}^{126} a_i x^i$  then  $R(x) = \sum_{i=0}^{63} a_i x^{2^{*i}}$ , which means that the attacker will easily get the values of  $a_i$  for  $i = 0..63$ . A simple solution to this problem is to use values of  $P(x)$  for which  $a_u = 0$  for  $u = 2..7$ .

A modification of the public-key system follows, which will make usage of any value of  $P(x)$  secure. We will describe the algorithm for  $n = 128$ , but it can be easily extended to be used for any value of  $n$ .

a) Public key encryption: any message  $m(x)$  of the length  $n$  as an input (plaintext) and evaluation of the polynomial  $P(m(x) * x^{128}) = c(x) \text{ mod } g(x)$  as an output (ciphertext). The evaluation operation will be implemented via a "White box" evaluation describe later.

b) Private key decryption: Given a ciphertext  $c'(x)$  calculate  $c(x)$ . Compute

$$\begin{aligned} P^{-1}(c(x)) * x^{2^{n-8}} &= P^{-1}(P(m(x) * x^{128}) * x^{2^{n-8}}) = \\ m(x) * x^{128} * x^{2^{n-8}} &= m(x) * x^{2^{n-1}} = m(x) \text{ mod } g(x). \end{aligned}$$

This is true because  $x^{2^{n-1}} = 1 \text{ mod } g(x)$ .

After the modification there are no terms with degrees of  $N < 128$  in the evaluation result of the polynomial  $P(m(x) * x^{128})$ , so there is no need of having an  $R(x)$  any more.

Polynomials  $L_0, L_1, \dots, L_{127}$  are generated, and values of  $B_N(x)$  are provided publicly for any  $N = 2^i(2k+1)$ , where  $i = 1, \dots, 128, r = 2k + 1, k = 0, 1, \dots, 127$ .

An encrypted message  $c'(x)$  then contains a 16 byte vector  $\sum B_N(x)$  and another 16 byte vector  $B = (b_0, b_1, \dots, b_{127})$ , where  $b_k = 0$  if the number of nonzero terms with the same value  $k$  in evaluation result is even and  $b_k = 1$  otherwise for  $N = 2^i(2k + 1), k = 0, \dots, 127$ .

Let  $P(x) = x^{2^{11}} + x^{2^{35}} + x^{2^{77}}$  and  $m(x) = x^3 + x^8$ . We have that

$$\begin{aligned} P(m(x) * x^{128}) &= (x^{131} + x^{136})^{2^{11}} + (x^{131} + x^{136})^{2^{35}} + (x^{131} + x^{136})^{2^{77}} \\ &= x^{131 \cdot 2^{11}} + x^{136 \cdot 2^{11}} + x^{131 \cdot 2^{35}} + x^{136 \cdot 2^{35}} + x^{131 \cdot 2^{77}} + x^{136 \cdot 2^{77}} \\ &= x^{131 \cdot 2^{11}} + x^{17 \cdot 2^{14}} + x^{131 \cdot 2^{35}} + x^{17 \cdot 2^{38}} + x^{131 \cdot 2^{77}} + x^{17 \cdot 2^{80}} \end{aligned}$$

We have that  $\sum B_N(x) = B_{131 \cdot 2^{11}}(x) \oplus B_{17 \cdot 2^{14}}(x) \oplus B_{131 \cdot 2^{35}}(x) \oplus B_{17 \cdot 2^{38}}(x) \oplus B_{131 \cdot 2^{77}}(x) \oplus B_{17 \cdot 2^{80}}(x)$  where  $B_N(x)$  are defined according to (0.1). Thus we have that for the vector  $B = (b_0, b_1, \dots, b_{63})$  in this case  $b_1 = 1$

Decoding procedure by the "owner" of the system will be as follows: based on the value  $B = (b_0, b_2, \dots, b_{127})$  and vector  $\sum B_N(x)$  the "owner" calculates:

$$\left( \sum B_N(x) \oplus \sum_{i=1}^{128} b_i \times L_i(x) \right) \times (L_0(x))^{-1} = c(x) \text{ mod } L(x). \quad (0.3)$$

After calculating  $c(x) = P(m(x) * x^{128})$  the "owner" of the system can decrypt the message:  $P^{-1}(c(x)) * x^{2^{n-8}} = m(x)$ , where  $P^{-1}$  is the compositional inverse of the polynomial  $P(x)$ .

### Implementation aspects and performance optimizations

Public key encryption of the proposed system requires evaluation of the polynomial  $P(m(x) \cdot x^{128})$  and then a modular reduction using white box implementation. The evaluation of  $P(m(x) \cdot x^{128})$  will require to count the values of  $m(x)^{2^i}$  for all  $i = 0..n$ , where  $2^n$  is the order of  $P(x)$ . This will require  $n$  squaring of polynomial  $m(x)$ . Let's denote by  $t$  the weight of  $P(x)$  and by  $s$  the weight of  $m(x)$ . Then the evaluation of  $P(m(x) \cdot x^{128})$  will have  $t * s$  terms, so  $t * s$  XORs of polynomials will be required to count  $\sum B_N(x)$ .

Calculation of  $c(x)$  from  $c'(x)$  will require 128 modulo two additions and one multiplication of polynomials as explained in section 3. Final decryption operation will require to compute  $P^{-1}(c(x)) * x^{2^{120}} = m(x)$ , where  $P^{-1}$  is a compositional inverse of polynomial  $P(x)$ . Calculation of  $P^{-1}(c(x))$  will require counting the values of  $c(x)^{2^i}$  for all  $i = 0..127$ , and XORing the resulting polynomials. If the weight of  $P^{-1}(x)$  is  $t$ , then  $t$  XORs will be required.

The memory required for storing the white box tables, I.E. the values of  $B_N(x)$  will be 16 bytes per value for each  $N = 2^i(2k + 1)$ , where  $i = 1, \dots, 128, r = 2k + 1, k = 0, 1, \dots, 127$ , resulting to overall  $128 \times 128 \times 16$  bytes = 256 Kbytes.

Some performance optimization were made to the described algorithm. One can notice, that for  $P(x) = \sum_{u=0}^n a_u x^{2^u}$  and  $m(x) = \sum_{v=0}^n a_v x^v$ ,

$$P(m(x)) = \sum_{v=0}^n a_v \times P(x^v). \quad (0.4)$$

In a similar way for  $m(x) = \sum_{l=0}^n a_l x^l$ ,

$$P^{-1}(c(x)) = \sum_{l=0}^n a_l \times P^{-1}(x^l). \quad (0.5)$$

So if we precalculate and store the values of  $\sum B_N(x)$  for terms in  $P(x^v)$  for each  $v = 0..127$  and the values of  $P^{-1}(x^l)$  for all  $l = 0..127$ , this will require  $128 * 16$  bytes of additional memory for each of encryption and decryption operations, but the performance will be increased dramatically. We also calculate the impact of  $P(x^v)$  on array B for each  $v = 0..127$ , which requires another  $128 * 16$  bytes of memory. After calculating and storing these values, instead of doing 128 squarings and  $t * s$  XORs for encryption, we'll do no squarings and just  $s$  XORs. In decryption we will skip doing the squarings. This optimizations speed up both encryption and decryption operations by about 2.2 times.

Performance tests were ran on Intel Core I5 CPU 1.6 GHZ processor, and the CryptoPP library's RSA-2048 implementation was used for comparison. A single encryption and decryption operations took 0.032ms and 0.041ms respectively for our algorithm, and 0.12ms and 5.46ms respectively for RSA-2048. So the algorithm described was 3.75x faster on encryption and 133x faster on decryption.

---

## Conclusion

---

In this paper a modification of white box encryption scheme [Khachatryan, Kyureghyan, 2015] based on permutation polynomials has been presented. Implementation aspects and performance optimizations were provided.

---

## Bibliography

---

- [Khachatryan, Kyureghyan, 2015] G. Khachatryan, M. Kyureghyan, Permutation polynomials and a new public key encryption - accepted for publication in Discrete Applied Mathematics journal- February 2015, 9 pages
- [Laigle, Chapuy, 2007] Y. Laigle-Chapuy, Permutation polynomials and applications to coding theory, Finite Fields, Appl.13, 58–70, 2007
- [Lidl, Niederreiter, 1983] R. Lidl, Niederreiter, Finite Fields, Addison Wesley, reading, MA, 1983.
- [Schwenk, Huber, 1998] J. Schwenk, K. Huber, Public key encryption and digital signatures based on permutation polynomials, Electron. Lett.34 (1998), 759–760.
- [Zeirler, 1959] N. Zeirler, Linear recurring sequences, J.Soc.Ind.Appl.Math.7,(1959), 31–48.
- [Diffie, Hellman, 1976] W. Diffie and M.E. Hellman, New Directions in Cryptography, IEEE Transactions on Information Theory, Vol. IT-22, Nov.1976, 644–654.
- [Rivest, Shamir, Adleman, 1978] R. Rivest, A. Shamir, L. Adleman, A Method for Obtaining Digital Signatures and Public-Key Cryptosystems, Communications of the ACM 21 (2), (1978), 120–126.
- [Miller, 1986] V.S.Miller, Use of Elliptic curves in cryptography, Advanced in Cryptology-Crypto-85 Proceedings, Springer-Verlag, (1986), 417–426.

---

**Authors' Information**

---



**Gurgen Khachatryan** *American University of Armenia*  
Yerevan, Armenia  
e-mail: [gurgenkh@aua.am](mailto:gurgenkh@aua.am)



**Martun Karapetyan** *Institute for Informatics and Automation Problems*  
*National Academy of Sciences of Armenia*  
Yerevan, Armenia  
e-mail: [martun.karapetyan@gmail.com](mailto:martun.karapetyan@gmail.com)

## CONVEXITY RELATED ISSUES FOR THE SET OF HYPERGRAPHIC SEQUENCES

Hasmik Sahakyan, Levon Aslanyan

**Abstract:** We consider  $D_m(n)$ , the set of all degree sequences of simple hypergraphs with  $n$  vertices and  $m$  hyperedges. We show that  $D_m(n)$ , which is a subset of the  $n$ -dimensional  $m + 1$ -valued grid  $\mathcal{E}_{m+1}^n$ , is not a convex subset of  $\mathcal{E}_{m+1}^n$ ; and give a characterization of the convex hull of  $D_m(n)$ .

**Keywords:** hypergraph, degree sequence, convexity.

**ACM Classification Keywords:** F.2.2: Nonnumerical Algorithms and Problems; G.2.2 Graph Theory

### Introduction

A hypergraph  $H$  is a pair  $(V, E)$ , where  $V$  is the vertex set of  $H$ , and  $E$ , the set of hyperedges, is a collection of non-empty subsets of  $V$ . The degree of a vertex  $v$  of  $H$ , denoted by  $d(v)$ , is the number of hyperedges in  $H$  containing  $v$ . A hypergraph  $H$  is simple if it has no repeated hyperedges. A hypergraph  $H$  is  $r$ -uniform if all hyperedges contain  $r$ -vertices. 2-uniform hypergraphs (edges contain exactly 2 vertices) are simply ordinary graphs.

Let  $V = \{v_1, \dots, v_n\}$ .  $d(H) = (d(v_1), \dots, d(v_n))$  is the degree sequence of hypergraph  $H$ . A sequence  $d = (d_1, \dots, d_n)$  is hypergraphic if there is a simple hypergraph  $H$  with degree sequence  $d$ . For a given  $m$ ,  $0 < m \leq 2^n$ , let  $H_m(n)$  denote the set of all simple hypergraphs  $([n], E)$ , where  $[n] = \{1, 2, \dots, n\}$ , and  $|E| = m$ , and  $D_m(n)$  denote the set of all hypergraphic sequences of hypergraphs in  $H_m(n)$ .

We investigate issues related to the characterization of the set of all hypergraphic sequences. The case of graphs is easy - a simple necessary and sufficient condition for the characterization of the set of degree sequences is known by the Erdos-Gallai Theorem [Erdos, Gallai, 1960], [Harary, 1969]:

**Theorem 1** (Erdos-Gallai) A decreasing sequence of non-negative integers  $(d_1, \dots, d_n)$  is the degree sequence for a simple graph if and only if:

$$\sum_{i=1}^n d_i \text{ is even;} \tag{1}$$

$$\sum_{i=1}^k d_i \leq k(k-1) + \sum_{i=k+1}^n \min\{k, d_i\} \text{ for } k = 1, \dots, n-1.$$

In general, the characterization of degree sequences for uniform hypergraphs is an open problem when  $r \geq 3$  (see [Berge, 1989], [Bill, 1988], [Bill, 1986], [BhanuSriv, 2002], [Colb, 1986], [KocayLi, 2007]).

The characterization of  $D_m(n)$ , which is not easier than the case of uniform hypergraphs, - is investigated in [Sah, 2009] - [Sah, 2015], [AslGroSahWag, 2015]. The problem has its interpretation in terms of multidimensional binary cubes; it is also known as a special case in discrete tomography problem, when an additional constraint/requirement – non-repetition of rows is imposed [SahAsl, 2010], [Sah, 2013]. Structures, properties, and several related partial results were obtained in [Sah, 2009] - [Sah, 2015] for  $D_m(n)$ . In this research we consider convexity issues related to the set  $D_m(n)$ .

Convex hull of degree sequences of  $k$ -uniform hypergraphs was investigated in [Koren, 1973],[BhanuSriv, 2002], [Klivans, Reiner, 2008], [Ricky Ini Liu, 2013]. It was shown by Koren [Koren, 1973] that the inequalities in (1) define a convex polytope  $D_n(2)$  of degree sequences of simple graphs, so that the sequences with even sum, lying in this polytope are exactly the degree sequences of the graphs on  $n$  vertices.

Analogous questions for  $k$ -uniform hypergraphs when  $k > 2$  investigated in [Klivans, Reiner, 2008], [Ricky Ini Liu, 2013]. Klivans and Reiner [Klivans, Reiner, 2008] verified computationally that the set of degree sequences for  $k$ -uniform hypergraphs is the intersection of a lattice and a convex polytope for  $k = 3$  and  $\leq 8$ . Ricky Ini Liu [Ricky Ini Liu, 2013] show that this does not hold for  $k \geq 3$  and  $n \geq k + 13$ .

In this paper we consider analogous convexity questions for  $D_m(n)$ .

---

### Structure of $D_m(n)$

---

Suppose that we consider the set of all hypergraphic sequences of hypergraphs  $([n], E)$ , and omit the restriction of non-repetition of hyperedges. Then, every integer sequence of length  $n$  with all component values between 0 and  $m$ , can serve as degree sequence of some hypergraph with the vertex set  $[n]$  and with  $m$  hyperedges.

Thus, the  $n$ -dimensional  $m + 1$ -valued integer grid  $\mathcal{E}_{m+1}^n$  of elements:  $\{(a_1, \dots, a_n) | 0 \leq a_i \leq m \text{ for all } i\}$  can be considered as the set of degree sequences of hypergraphs with the vertex set  $[n]$  and with  $m$  hyperedges; and in this manner,  $D_m(n) \subseteq \mathcal{E}_{m+1}^n$ .

In this section we consider the structure of  $D_m(n)$  in  $\mathcal{E}_{m+1}^n$ .

Component-wise partial order is defined on  $\mathcal{E}_{m+1}^n$ :  $(a_1, \dots, a_n) \leq (b_1, \dots, b_n)$  if and only if  $a_i \leq b_i$  for all  $i$ , and  $r(a_1, \dots, a_n) = a_1 + \dots + a_n$  is the rank of an element  $(a_1, \dots, a_n)$ . An illustration of  $\mathcal{E}_{m+1}^n$  can be given by the Hasse diagram. Figure1 illustrates the Hasse diagram of  $\mathcal{E}_5^3$ .



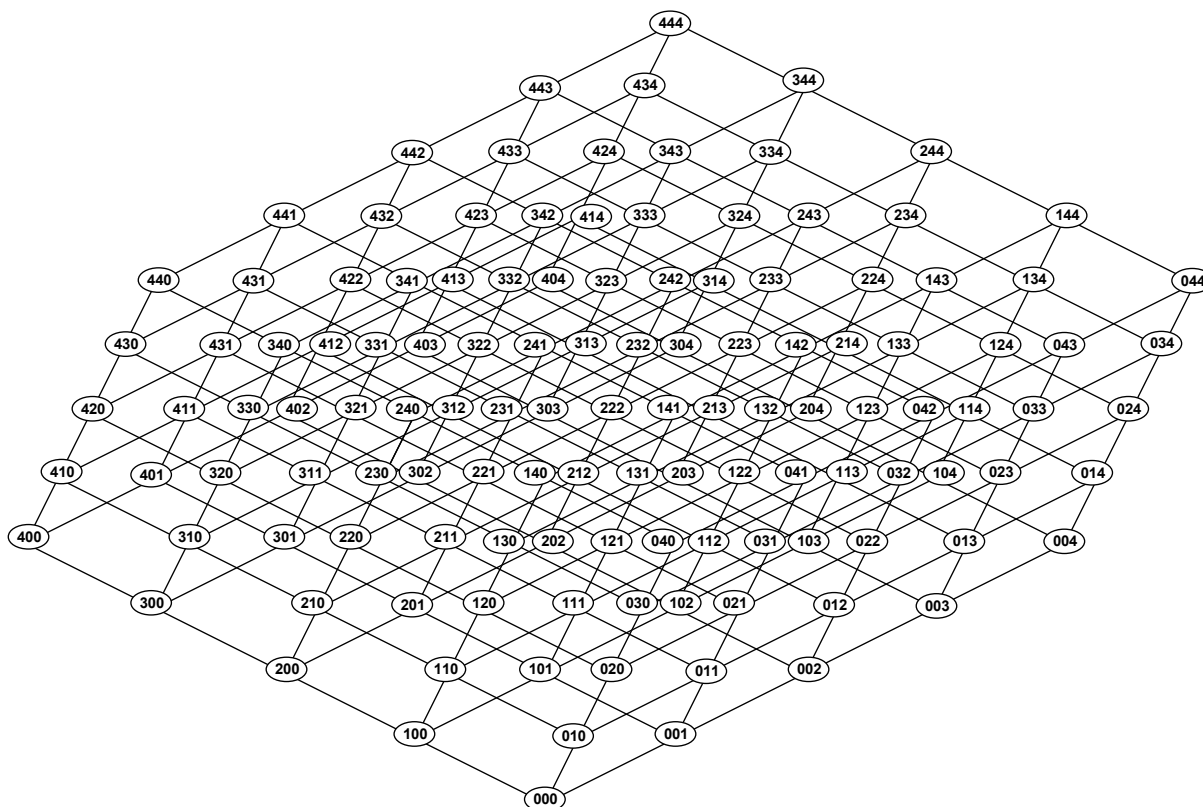


Figure 1. The Hasse diagram of  $E_5^3$

**Opposite elements in  $E_{m+1}^n$**

A pair of elements  $d, \bar{d} \in E_{m+1}^n$  are called opposite if one can be obtained from the other by inversions of component values: if  $d = (d_1, \dots, d_n)$ , then  $\bar{d} = (m - d_1, \dots, m - d_n)$ .

**Boundary elements of  $D_m(n)$ .**

We call  $(d_1, \dots, d_n) \in D_m(n)$  an *upper boundary /lower boundary/ element* of  $D_m(n)$  if no  $(a_1, \dots, a_n) \in E_{m+1}^n$  with  $(a_1, \dots, a_n) > (d_1, \dots, d_n)$  / with  $(a_1, \dots, a_n) < (d_1, \dots, d_n)$  / belongs to  $D_m(n)$ .

Let  $\hat{D}_{max}$  and  $\check{D}_{min}$  denote the sets of upper and lower boundary elements of  $D_m(n)$ , respectively.

**Interval in  $E_{m+1}^n$ .**

For a pair of elements  $d', d'', d' \leq d''$  of  $E_{m+1}^n$ ,  $E(d', d'')$  denotes the minimal subgrid/interval in  $E_{m+1}^n$  spanned by these elements:  $E(d', d'') = \{a \in E_{m+1}^n \mid d' \leq a \leq d''\}$ .

**Theorem 2** ([Sah, 2009]).  $D_m(n)$  is a union of intervals spanned by the pairs of opposite elements of  $\hat{D}_{max}$  and  $\check{D}_{min}$ :

$$D_m(n) = \bigcup_{\hat{D} \in \hat{D}_{max}, \check{D} \in \check{D}_{min}} E(\check{D}, \hat{D}),$$

where  $(\hat{D}, \check{D})$  are pairs of opposite elements.

An illustration is given in Figure 2 by the example of  $D_4(3)$  in  $\Xi_5^3$ :

$$\widehat{D}_{max} = \{(3,3,3), (4,2,2), (2,4,2), (2,2,4)\},$$

$$\check{D}_{min} = \{(1,1,1), (0,2,2), (2,0,4), (2,2,0)\},$$

$$D_4(3) =$$

$$E((1,1,1), (3,3,3)) \cup E((0,2,2), (4,2,4)) \cup E((2,2,0), (2,2,4)) \cup E((2,0,2), (2,4,2)).$$

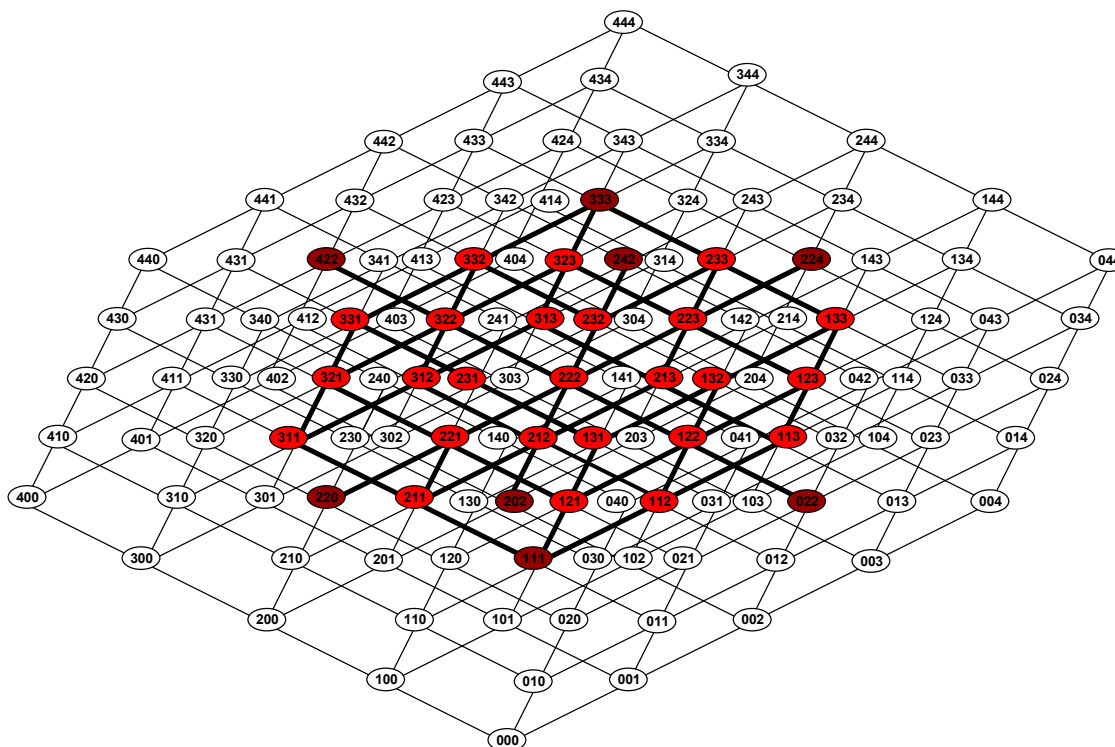


Figure 2

$D_4(3)$  in  $\Xi_5^3$ : vertices in red compose  $D_4(3)$ , and vertices in darker red compose sets  $\widehat{D}_{max}$  and  $\check{D}_{min}$ .

### Non-convexity of $D_m(n)$ in $\Xi_{m+1}^n$

In this section we show that  $D_m(n)$  is not a convex set in  $\Xi_{m+1}^n$ .

**Convex set.** [Birkhoff, 1948] A subset  $S$  of the poset  $P$  is *convex* whenever  $a \in S, b \in S$  and  $a \leq b$  imply  $[a, b] \in S$ .

It follows from the definition that each interval  $E(\check{D}, \widehat{D})$  spanned by opposite boundary elements is a convex set in  $\Xi_{m+1}^n$ .

Nevertheless we prove that  $D_m(n)$  being a union of convex sets, - is not convex.

**Theorem 3.**  $D_m(n)$  is not convex in  $\mathcal{E}_{m+1}^n$ , when  $1 < m < 2^n - 1$ .

We omit the details of the proof and just bring the outline. First we show that  $D_m(n)$  is convex for the following values of  $m$ :

- a)  $m = 1$ . We show that  $D_m(n) = E(\tilde{0}, \tilde{m})$ , which coincides with  $\mathcal{E}_{m+1}^n$ , and thus, is a convex set.
- b)  $m = 2^n$ . In this case  $D_m(n) = E((2^{n-1}, \dots, 2^{n-1}), (2^{n-1}, \dots, 2^{n-1}))$  – that is 1 point of  $\mathcal{E}_{m+1}^n$ .
- c)  $m = 2^n - 1$ . Here  $D_m(n) = E((2^{n-1} - 1, \dots, 2^{n-1} - 1), (2^{n-1}, \dots, 2^{n-1}))$  – this is an interval of  $\mathcal{E}_{m+1}^n$ , and thus, is a convex set.

Then we prove that for the following cases:

- d)  $1 < m \leq 2^{n-1}$
- e)  $2^{n-1} < m < 2^n - 1$

there always exist two comparable elements  $a < b$  in  $D_m(n)$ , such that the spanned interval  $E(a, b)$  in  $\mathcal{E}_{m+1}^n$  contain an element  $c \notin D_m(n)$ .

Consider an example in Figure 3.

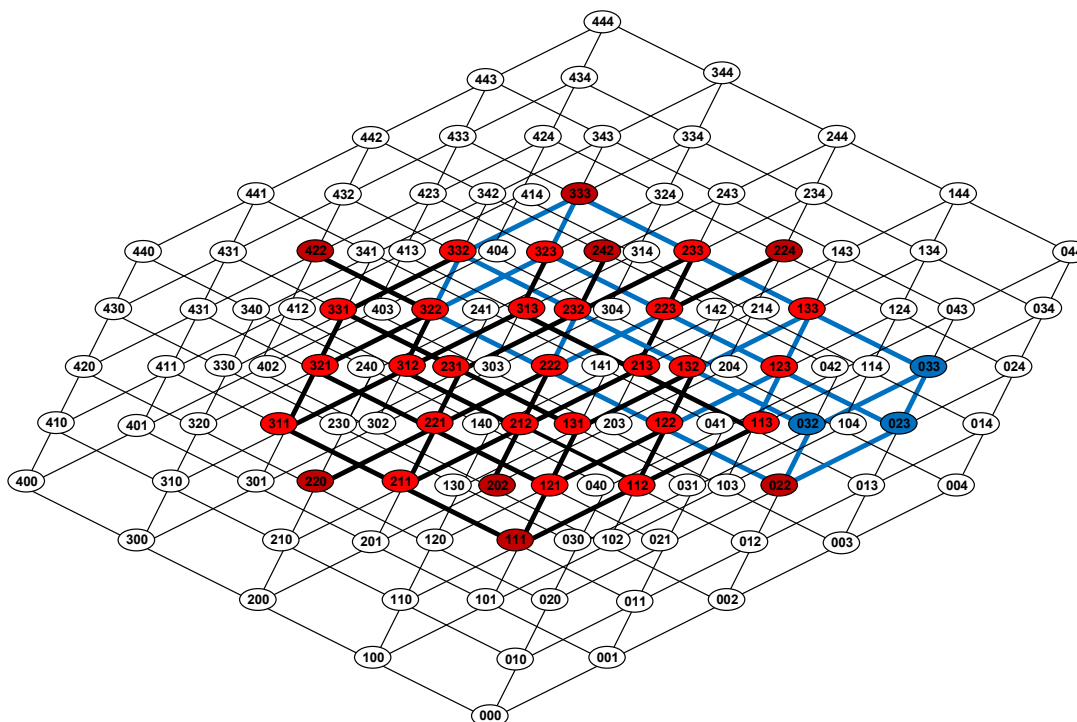


Figure 3

The elements  $(0,2,2)$  and  $(3,3,3)$  belong to  $D_4(3)$ , and  $(0,2,2) < (3,3,3)$ . However the elements  $(0,3,2), (0,2,3), (0,3,3)$  of  $\mathcal{E}_5^4$ , which are greater than  $(0,2,2)$ , and less than  $(3,3,3)$ , - do not belong to  $D_4(3)$ .

**Convex hull of  $D_m(n)$**

In this section we characterize the convex hull of  $D_m(n)$ .

**Convex hull** ([Eggleston, 1958])

Let  $S$  be a nonempty subset of  $R^n$ . Then among all convex sets containing  $S$  (these sets exist, e.g.,  $R^n$  itself) there exists the smallest one, namely, the intersection of all convex sets containing  $S$ .

This set is called the *convex hull of  $S$*  (denote by:  $Conv(S)$ ).

In our case we consider the intersection of  $Conv(D_m(n))$  and  $Z^n$  - in other words we consider the integer points of  $Conv(D_m(n))$ .

Notice that  $\mathcal{E}_{m+1}^n$  itself corresponds to some convex set of  $R^n$ .  $D_m(n) \subseteq \mathcal{E}_{m+1}^n$  is also contained in the mentioned convex set. We are interested in finding the smallest convex subset of  $\mathcal{E}_{m+1}^n$ , containing  $D_m(n)$ . We denote this set by  $C_{D_m(n)}$ .

**Theorem 4.**  $C_{D_m(n)} = \bigcup_{\widehat{D} \in \widehat{D}_{max}, \check{D} \in \check{D}_{min}} E(\check{D}, \widehat{D})$  (the union is by all pairs  $(\widehat{D}, \check{D})$  and not only by opposite pairs).

We prove the theorem by showing first that the set  $\bigcup_{\widehat{D} \in \widehat{D}_{max}, \check{D} \in \check{D}_{min}} E(\check{D}, \widehat{D})$  is a convex set, and then - that this is the smallest convex set containing  $D_m(n)$ .

An illustration is in Figure 4.

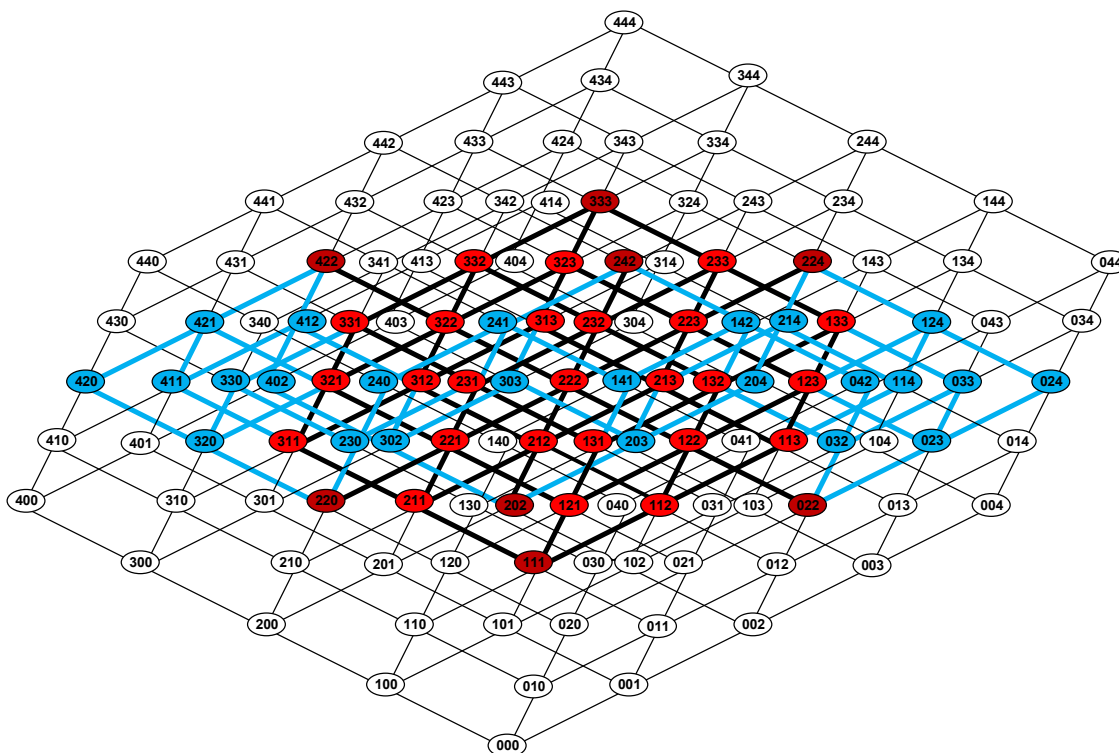


Figure 4

$C_{D_4(3)}$  in  $\Xi_5^3$ , where the elements of  $D_4(3)$  are in red color, and the elements of  $C_{D_4(3)}$  are in red and blue colors.

### Corollary.

- The smallest convex subset of  $\Xi_{m+1}^n$  containing  $D_m(n)$  is the convex hull of the set  $(\widehat{D}_{max} \cup \check{D}_{min})$ .
- Each element  $d$  of  $(\widehat{D}_{max} \cup \check{D}_{min})$  is an extreme point of  $C_{D_m(n)}$  since  $\bigcup_{\widehat{D} \in \widehat{D}_{max}, \check{D} \in \check{D}_{min}} E(\check{D}, \widehat{D}) \setminus \{d\}$  is a convex set.

### Conclusion

We considered  $D_m(n)$ , the set of all degree sequences of hypergraphs with  $n$  vertices and  $m$  hyperedges, as a subset of the  $n$ -dimensional  $m + 1$ -valued grid  $\Xi_{m+1}^n$ . We showed that  $D_m(n)$  is not a convex subset of  $\Xi_{m+1}^n$ , and characterized the convex hull of  $D_m(n)$ .

### Bibliography

- [AslGroSahWag, 2015] Levon Aslanyan, Hans-Dietrich Gronau, Hasmik Sahakyan, Peter Wagner, Constraint Satisfaction Problems on Specific Subsets of the  $n$ -Dimensional Unit Cube, CSIT 2015, Revised Selected Papers, IEEE conference proceedings, p.47-52, DOI:10.1109/CSITechnol.2015.7358249
- [Berge, 1989] Berge C., Hypergraphs: Combinatorics of Finite Sets, North-Holland, 1989
- [BhanuSriv, 2002] Bhanu Murthy N.L., Murali K. Srinivasan, The polytope of degree sequences of hypergraphs, Linear Algebra Appl. 350 (2002) 147–170
- [Bill, 1986] Billington D., Lattices and Degree Sequences of Uniform Hypergraphs. Ars Combinatoria, 21A, 1986, 9-19.
- [Bill, 1988] Billington D., Conditions for degree sequences to be realisable by 3-uniform hypergraphs”. The Journal of Combinatorial Mathematics and Combinatorial Computing”. 3, 1988, 71-91.
- [Birkhoff, 1948] G. Birkhoff, Lattice Theory. American Mathematical Society Colloquium Publications, Volume XXV. American Mathematical Society, 1948.

- [Colb, 1986] Colbourn Charles J., Kocay W.L. and Stinson D.R., Some NP-complete problems for hypergraph degree sequences. *Discrete Applied Mathematics* 14, p. 239-254 (1986))
- [Eggleston, 1958] H. G. Eggleston, Chapter 1 - GENERAL PROPERTIES OF CONVEX SETS , pp. 1-32, Publisher: Cambridge University Press, 1958 Online Publication, 2010, DOI: <http://dx.doi.org/10.1017/CBO9780511566172.002>
- [Erdos,Gallai, 1960] P. Erdos and T. Gallai. Graphs with given degrees of vertices. *Mat. Lapok*, 11 (1960), 264-274.
- [Harray, 1969] F. Harary. *Graph Theory*. Addison Wesley, Reading, 1969.
- [Klivans, Reiner, 2008] C. Klivans and V. Reiner, Shifted set families, degree sequences, and plethysm. *Electron. J. Combin.*, 15(1):Research Paper 14, 35, 2008.
- [KocayLi, 2007] Kocay William and Li Pak Ching, On 3-hypergraphs with equal degree sequences, *Ars Combin.* 82 (2007), 145–157.
- [Koren, 1973] Michael Koren, Extreme degree sequences of simple graphs. *J. Combinatorial Theory Ser. B*, 15:213–224, 1973.
- [Ricky Ini Liu, 2013] Ricky Ini Liu, Nonconvexity of the set of hypergraph degree sequences, *Electronic journal of combinatorics* 20(1) (2013), #P21.
- [Sah, 2009] H. Sahakyan, Numerical characterization of n-cube subset partitioning, *Discrete Applied Mathematics*, 157 (2009), pp. 2191-2197.
- [Sah, 2013] Sahakyan Hasmik, “(0,1)-matrices with different rows”, Ninth International Conference on Computer Science and Information Technologies, Revised Selected Papers, IEEE conference proceedings, 2013.
- [Sah, 2014] Sahakyan H., Essential points of the n-cube subset partitioning characterization, *Discrete Applied Mathematics*, vol. 163, part 2, 2014, pp. 205-213
- [Sah, 2015] Sahakyan H., On the set of simple hypergraph degree sequences, *Applied Mathematical Sciences*, v. 9, 2015, no. 5, pp. 243-253, Hikari ltd
- [SahAsl, 2010] Hasmik Sahakyan, Levon Aslanyan, Linear program form for ray different discrete tomography, *International Journal “Information Technologies and Knowledge”*, Vol. 4, Number 1, 2010, p.41-50.

---

## Authors' Information

---



**Hasmik Sahakyan** – Institute for Informatics and Automation Problems of the National Academy of Sciences of Armenia, scientific secretary; 1 P. Sevak str., Yerevan 0014, Armenia; e-mail: [hsahakyan@sci.am](mailto:hsahakyan@sci.am)

*Major Fields of Scientific Research: Combinatorics, Discrete Tomography, Data Mining.*



**Levon Aslanyan** – Institute for Informatics and Automation Problems of the National Academy of Sciences of Armenia, head of department; 1 P. Sevak str., Yerevan 0014, Armenia; e-mail: [lasl@sci.am](mailto:lasl@sci.am)

*Major Fields of Scientific Research: Discrete analysis – algorithms and optimization, pattern recognition theory, information technologies.*

## REPRESENTING STRATEGIC ORGANIZATIONAL KNOWLEDGE VIA DIAGRAMS, MATRICES AND ONTOLOGIES

Dmitry Kudryavtsev, Anna Menshikova, Tatiana Gavrilova

**Abstract:** *The paper describes methods and tools for organizational knowledge representation in the field of strategic management. Visual and matrix/table-based methods are actively used for knowledge representation in this domain. Diagrams solve problems associated with the managerial thinking (cognitive challenges), managerial communications and coordination (social problems), and the ability of managers to motivate and involve their employees (emotional problems). On the other hand, there are types of information and tasks that are better supported by matrices. In order to effectively combine diagrams with matrices the paper suggests multi-representation of organizational knowledge using ontologies. Such multi-representation capabilities for organizational knowledge are already supported by some enterprise architecture management software tools. Two of these tools are described in the paper.*

**Keywords:** *organizational knowledge, knowledge management methods, knowledge representation, knowledge structuring, ontologies, strategic management.*

**ACM Classification Keywords:** *A.0 General Literature - Conference proceedings*

---

### Introduction

---

Nowadays there are more and more new methods and tools for the effective operation of corporate knowledge. This article describes some results of the project INNOVARRA "Innovations in Organizational Knowledge Management: Typology, Methodology and Recommendations", aims to identify and develop knowledge management (KM) methods and tools, which are the most appropriate for particular knowledge type and domain of the company. Various enterprise knowledge domains (e.g. product/service knowledge, customer knowledge, operations management or strategic management knowledge etc.) have different knowledge characteristics and knowledge types. Systematization of knowledge types, characteristics and domains in INNOVARRA project is designed to differentiate KM methods and tools better suited for a particular knowledge domain. The project is based on the idea of the “triad” (Figure 1): “knowledge domain – the type or characteristics of knowledge – a method or a tool of knowledge management”. Analysis and systematization of studies linking types and domains of expertise with KM methods and tools was performed in the INNOVARRA project in order to differentiate KM methods and tools. Three tracks of the INNOVARRA project are considering the examples of KM methods and tools in several areas, namely: management of customer knowledge and knowledge about



the products and / or services, knowledge in the field of operations management, knowledge in the field of strategic management and organizational development. This paper describes preliminary results for knowledge representation methods in the field of strategic management.

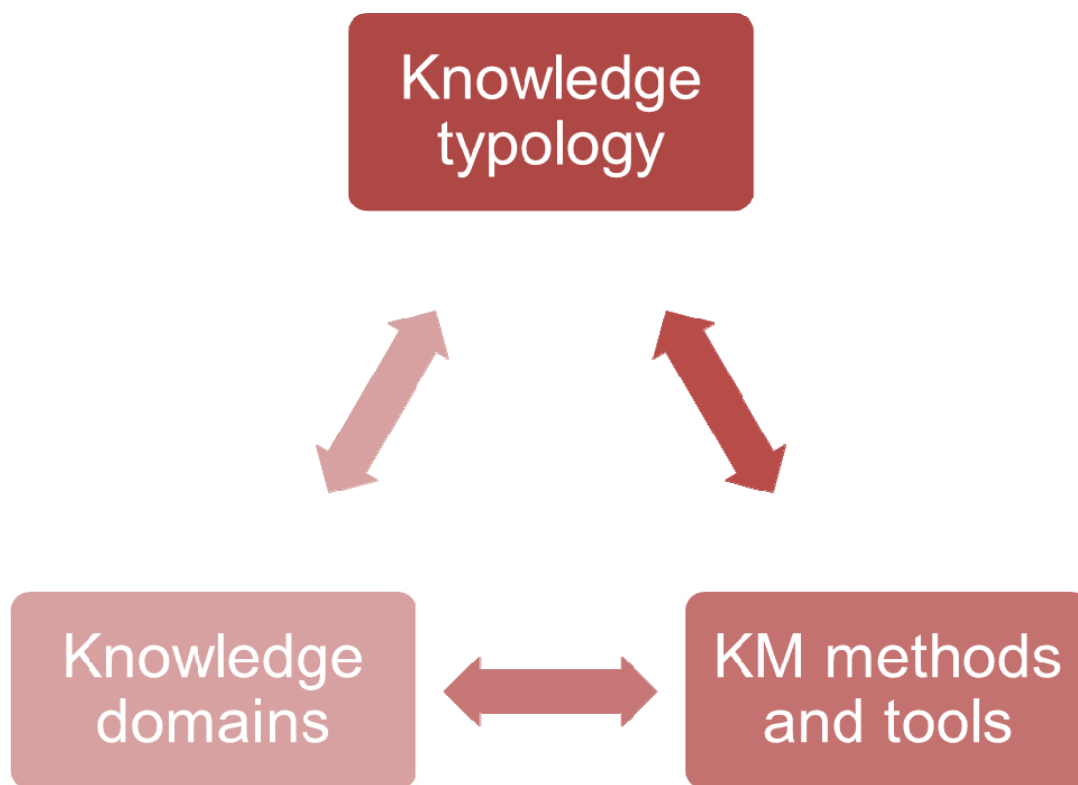


Figure 1. “Triad” of differentiation of KM methods and tools

An analysis of the literature identified the following problems in the development and implementation of the strategy: 1) cognitive (information overload, the rigidity of the old points of view, etc.) 2) social (different viewpoints of the team members and the need for their integration, etc.) 3) emotional (lack of a sense of belonging to strategies, etc.). Diagrams and matrix/table-based methods helps to overcome these problems, but should be used cooperatively. This cooperation can be provided by ontology-based tools. In the strategic management domain this technology can be implemented through ontology-based enterprise architecture management tools, which are initially aimed to process knowledge about an organization.

---

### **The use of visual methods in strategic management**

---

Strategic planning processes are one of the most difficult issues which managers face today. It can be an overwhelming challenge – at the same time to take into account the development of new technologies and social trends, and the behavior of competitors, customers and regulatory authorities,

changes in the legal, environmental and financial base. The problem is compounded by the limited time, market uncertainty, the constant changes and internal tensions. Strategic planning task is further complicated by the need for communication, implementation and monitoring of these decisions. Taken together, these activities create numerous cognitive (eg. information overload), social (eg. coordination of several groups and hierarchical levels) and emotional (eg. employee involvement) problems for the manager. However, visualization – a graphical representation of data, information and knowledge – may offer significant advantages in each of these three spheres. And, together with tables and diagrams, it is becoming more and more popular knowledge management tool in the domain of strategic management. The relationship between the key issues of the process of strategic management and the benefits that have been made possible thanks to the visualization are shown in Table 1. This table shows the potential benefits from the use of graphic representation of strategic content for the strategy development process.

Table 1. Strategic issues and ways to solve them with the help of visualization [Eppler & Platts, 2009]

<b>Possible problems in strategy development</b>	<b>Corresponding strengths of visualization</b>
<b>Cognitive</b>	
<b>Information overload</b> Overload due to the large amount of information in the analysis, its complexity	<b>Facilitating elicitation and synthesis of information</b> The visual channel improves the perception, there is compression of information, the patterns and structures of data set are seen more clearly
<b>Stuck in old view points</b> The development of strategic options often requires novel perspectives and divergent thinking	Visual methods enable reframing, change points of view, inspire creativity and contribute to perspective switching
<b>Biased comparison and evaluation</b>	Better, more exhaustive comparisons

<p><b>Paralysis by analysis</b></p> <p>Omission of strategic information due to the large flow of information relating to daily operations</p>	<p>Visualization helps to remember the current strategic conversations, visual recall is better than verbal recall</p>
<p><b>Social</b></p>	
<p><b>Diverging views or assumptions between team members</b></p> <p>Strategy development and formulation requires collective sense making processes and input from various team members</p>	<p><b>Integrating different perspectives</b></p> <p>Visualization can equilibrate participation and reduce the dominance of certain participants, identifies areas of disagreement</p>
<p><b>Incomplete communication of basic assumptions</b></p> <p>Managers need to assure that their reasoning is properly understood by employees</p>	<p><b>Assisting mutual understanding</b></p> <p>Visual tools often make basic assumptions explicit</p>
<p><b>Coordination difficulties</b></p> <p>Strategizing requires coordination both in communications and actions. This is especially true for globally dispersed teams</p>	<p>Visual artefacts provide explicit reference points for mutual coordination and alignment, the ability to share network modeling</p>
<p><b>Emotional</b></p>	
<p><b>Lacking identification with strategy</b></p>	<p><b>Creating involvement and engagement</b></p> <p>Pictures can create involvement and engage people’s imagination</p>
<p>Employees should perceive the strategy as something worthwhile pursuing, something that aspires and motivates</p>	<p>Helps to inspire and motivate people</p>
<p>The strategy needs to be communicated to employees convincingly</p>	<p>Visualization is ideally suited for convincing communication and presentation purposes</p>

Cognitive, social and emotional benefits could be better represented by a different genre of strategy visualization with the help of typical visualization tools and formats used for the strategy process as shown in Table 2.

Table 2. Four genres of strategy visualization methods [Eppler & Platts, 2009]

<b>Visualization Method Type</b>	<b>Main Features</b>	<b>Examples of Typical Visual Formats</b>
<b>Structuring Methods</b> (Analysis Phase)	Provide a ready-to-use structure (incl. categories) to organize and synthesize information	Bar diagram, line chart, system/loop diagram, 2by2 positioning matrices (BCG, McKinsey, SWOT), Porter's five forces diagram, S-curve diagram strategy chart, product-market diagram
<b>Elaboration Methods</b> (Development Phase)	Provide rules and a relatively open structure to elaborate on information, discover new patterns, build a common understanding and develop options	Decision tree, Ansoff matrix, morphological box, knowledge map, concept map, Mind Map, Parameter Ruler, influence diagrams, strategy canvas
<b>Sequencing Methods</b> (Planning Phase)	Provide rules, categories and graphic structures to organize information, such as tasks or goals, chronologically to prepare action	Timeline, flowchart, Gantt chart, road mapping, CPM diagram (critical path method), PERT diagram, swim lane diagram, loop diagram, Synergy Map
<b>Interaction Methods</b> (Implementation Phase)	Provide an interface to capture, aggregate, present and explore information.	Management controlling dashboard/cockpit, Strategy Map, visual metaphors, tracking diagrams such as flight plans

Cognitive advantages of visual representations include facilitating the identification and compilation of information. This provides new perspectives that allow carrying out comprehensive comparison of alternatives and facilitate planning of the sequence of actions. Social benefits include different integrations, helping people to understand and support each other. Finally, the emotional benefits include creating a sense of participation and involvement, providing inspiration and forging closer ties.

Additional review and classifications of visual knowledge processing techniques can be found in [Gavrilova, Gulyakina, 2011; Kudryavtsev, Gavrilova, 2016].

---

### **The use of table / matrix methods of knowledge structuring in strategic management**

---

With the increasing complexity of management technology, demands on decision support tools and methods to solve specific business problems are growing as well. It is important that the tools, processes and structures supporting management technologies have the following characteristics:

- reliability;
- economic and practical feasibility (would not be too complex or resource intensive);
- ability to integrate (would work together with other frames, processes and tools already deployed in business);
- flexibility (the ability to adapt to the specific context of business goals, market environment, the available resources and information, corporate culture, etc.).

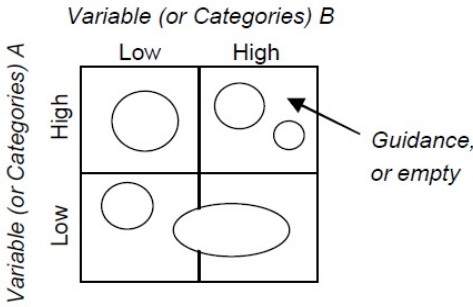
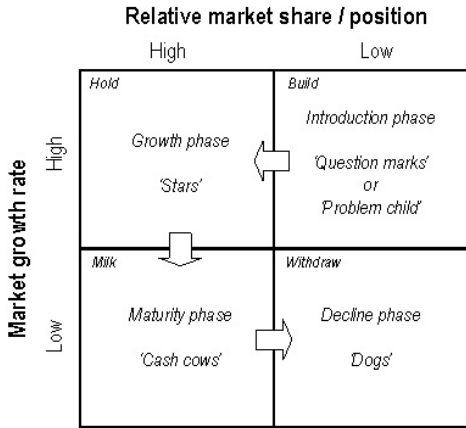
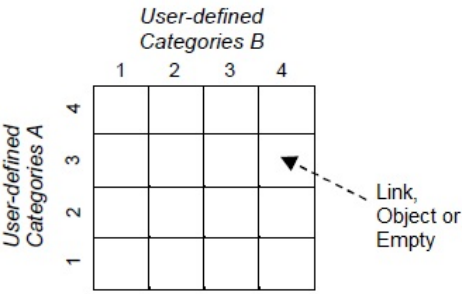
Such tools can take many forms, including matrix or state-space solutions, matrix connections, tables, profiles, checklists, taxonomy, software, and combinations thereof.

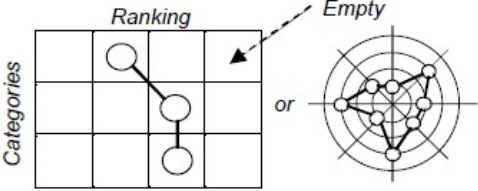
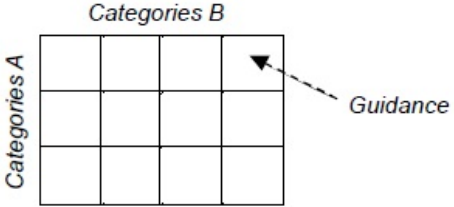

As noted above, there are many different types of control instruments, methods, processes, structures and patterns. In this paper, the focus is just on the "Matrix Tools" presented in Table 3. Such a structure can facilitate understanding and help in decision-making, or recommend specific managerial actions.

Classical “2x2 matrix” is a typical example of this type of a tool, widely used by consultants and managers in business, as well as by researchers. These tools are widely discussed in the literature, although a guide on applying them in practice could be rarely found. Phaal with the colleagues [Phaal et al, 2006] offered 4 different types of matrices: a matrix or state-space solutions, matrix connections, tables, profiles; on the basis of a large sample of more than 850 tools and mechanisms of matrix tools.

Matrix methods are relatively simple and orthogonal. Most often, these structures are two-dimensional, but the dimension may be higher. The structure connects the key aspects of the problems to be solved. The axes can be divided into categories or specific variables that can be qualitative and quantitative, discrete and continuous. The matrix can either already contain text, structured along the axes and related categories, or can be “empty”, that allows the user to explore the relative position of the various options for the relationship between the key dimensions and categories.

Table 3. Matrix tools (revised and extended version of [Phaal et al, 2006])

Types	Descriptions and examples																																			
<p>Positioning and decision matrices</p> 	<p>Categories are usually split into multiple values. If the matrix is empty, attention should be focused on the study of the mutual arrangement of the various options. This is the most common type of instrument.</p> <p><b>Example: BCG Matrix</b></p> 																																			
<p>Matrices of relationships (Generic grids)</p> 	<p>Axes are divided into a plurality of different categories, the number and content of which are specified by the user. The matrix provides a structure that allows the user to explore the relationship between the axes and associated with them categories.</p> <p><b>Example: Responsibility Matrix</b></p> <table border="1" data-bbox="730 1541 1252 1765"> <thead> <tr> <th></th> <th>ROLE 1</th> <th>ROLE 2</th> <th>ROLE 3</th> <th>ROLE 4</th> </tr> </thead> <tbody> <tr> <th>TASK 1</th> <td>R</td> <td>C</td> <td>I</td> <td>A</td> </tr> <tr> <th>TASK 2</th> <td>I</td> <td>I</td> <td>R</td> <td>A</td> </tr> <tr> <th>TASK 3</th> <td>C</td> <td>R</td> <td>A</td> <td>I</td> </tr> <tr> <th>TASK 4</th> <td>A</td> <td>R</td> <td>I</td> <td></td> </tr> <tr> <th>TASK 5</th> <td>R</td> <td>A</td> <td>C</td> <td>I</td> </tr> <tr> <th>TASK 6</th> <td>C</td> <td>C</td> <td>A+R</td> <td>I</td> </tr> </tbody> </table> <p>R=responsible A=accountable C=consulted I=informed</p>		ROLE 1	ROLE 2	ROLE 3	ROLE 4	TASK 1	R	C	I	A	TASK 2	I	I	R	A	TASK 3	C	R	A	I	TASK 4	A	R	I		TASK 5	R	A	C	I	TASK 6	C	C	A+R	I
	ROLE 1	ROLE 2	ROLE 3	ROLE 4																																
TASK 1	R	C	I	A																																
TASK 2	I	I	R	A																																
TASK 3	C	R	A	I																																
TASK 4	A	R	I																																	
TASK 5	R	A	C	I																																
TASK 6	C	C	A+R	I																																

Types	Descriptions and examples																			
<p>Generic scored profile</p> 	<p>One axis is divided into separate defined categories, and the other indicates the scale that allows the user to evaluate an action in terms of specific categories. The tool may be in the form of radial graph.</p>																			
<p>Generic table</p> 	<p>The axes are divided into individual, specific, pre-defined categories. The matrix typically contains text providing information about the axes and associated categories.</p> <p>Example:</p> <table border="1" data-bbox="775 891 1393 1227"> <thead> <tr> <th colspan="2" rowspan="2"></th> <th colspan="2">Strategy</th> </tr> <tr> <th>Corporate strategy</th> <th>R&amp;D strategy</th> </tr> </thead> <tbody> <tr> <th rowspan="4">Area of influence</th> <th>Resources</th> <td>Allocation between functions - marketing, production, R&amp;D, ...</td> <td>Allocation between projects</td> </tr> <tr> <th>Objectives</th> <td>Related to business environment</td> <td>Related to corporate environment</td> </tr> <tr> <th>Business areas</th> <td>Product / market strategy, Product / market mix</td> <td>Technology / product strategy, Portfolio balance</td> </tr> <tr> <th>Timescale</th> <td>Balance between long / medium / short term</td> <td>Balance between long / medium / short term</td> </tr> </tbody> </table>			Strategy		Corporate strategy	R&D strategy	Area of influence	Resources	Allocation between functions - marketing, production, R&D, ...	Allocation between projects	Objectives	Related to business environment	Related to corporate environment	Business areas	Product / market strategy, Product / market mix	Technology / product strategy, Portfolio balance	Timescale	Balance between long / medium / short term	Balance between long / medium / short term
				Strategy																
		Corporate strategy	R&D strategy																	
Area of influence	Resources	Allocation between functions - marketing, production, R&D, ...	Allocation between projects																	
	Objectives	Related to business environment	Related to corporate environment																	
	Business areas	Product / market strategy, Product / market mix	Technology / product strategy, Portfolio balance																	
	Timescale	Balance between long / medium / short term	Balance between long / medium / short term																	
<p>Canvas (table-based template)</p> 	<p>Example: Business Model Canvas [Osterwalder, Pigneur, 2010]</p> <p>The tool is used as a visual chart with elements of business models describing a firm's or product's value proposition, infrastructure, customers, and finances. It allows aligning and challenging activities of the companies.</p>																			

The following benefits of matrix methods are allocated in the studies:

1. They are relatively simple, both in terms of their concept and in use. Most of these tools can be represented as a simple scheme.
2. Typically tools based on matrices are flexible – they can be applied to specific situations in the company. You may need adjustment in accordance with the current context, which in general meets the criteria of flexibility.

3. Assuming that axis and parameters can be combined, matrix-based tools have the ability to be linked to form a more powerful integrated set of tools.

However, matrix tools have potential drawbacks:

1. Many practical problems or issues cannot be simplified to two dimensions that make the matrix tools ignore other important factors.
2. The use of these instruments as a rule requires research or settings, which may not be an easy task.
3. The use of tools of this class is impossible when the theoretical foundations of the instrument are not clear, or if knowledge and skills necessary for their effective application are inadequate.

---

### **Combining diagrams and matrices**

---

The effects of tables and graphs on elementary tasks are generally well studied [DeSanctis & Jarvenpaa, 1985; Jarvenpaa, 1989; Jarvenpaa & Dickson, 1988; Vessey, 1991]. Specifically, Jarvenpaa and Dickson summarized several studies, finding that graphs lead to faster or better performance for most elementary tasks, including summarizing data, showing trends, comparing points and patterns, and showing deviations. Tables, however, lead to better performance for the task of reading the value of single points. Two theories serve to predict the effects of tables and maps on problem solving, and both preach compatibility between the demands of the task and the representation of information. The first, cognitive fit [Vessey, 1991], recommends spatial information for spatial tasks and symbolic information for symbolic tasks, and explains the effects found by Jarvenpaa and Dickson [Jarvenpaa, Dickson, 1988] in that light. The second theory, the Proximity Compatibility Principle (PCP), espouses physical proximity of data if the task demands its integration [Wickens, 1992; Wickens, Merwin, & Lin, 1994].

There are two papers [Ghoniem et al, 2005; Keller et al, 2006] that compared the representation power of matrix and node-link diagrams. Ghoniem et al. [Ghoniem et al. 2005] showed that matrices outperform node-link diagrams for large or dense graphs in several low-level reading tasks, except path finding. This difference is supported by user study experiments conducted by Keller et al., as they found that node-link diagrams offer better visual representation for small, uncomplicated entities, but they are a complex form of representation for large systems and propagation across systems [Keller et al, 2006].

So diagrams and matrices/tables have their own advantages and disadvantages. It is important to find cognitive fit between task, type of information and representation format. Typically the same information can be represented in either format for different purposes. Let's consider a couple of strategic management models, which are represented in both visual and table form.

The first example of using visual forms for strategic knowledge representation was described in [Gavrilova, Alsufiev, Yanson, 2014], where visual conversion of classical business model CANVAS [Osterwalder et al, 2005] into a mind map was proposed (see Fig. 2). That map suggests the most



compact and compressed form of strategic company knowledge. The proposed business model template in the form of a mind map uses a blend of modern theories of knowledge engineering, cognitive sciences, and Gestalt psychology. The presented approach employs the building graphs methods and techniques, particularly mind maps [Buzan, 2003]. The canvas business model traditionally consists of nine blocks that reflect the structure of business processes: key partners, key resources, key activities, value proposition, sales channels, customer segment, customer relationships, revenue streams, costs.

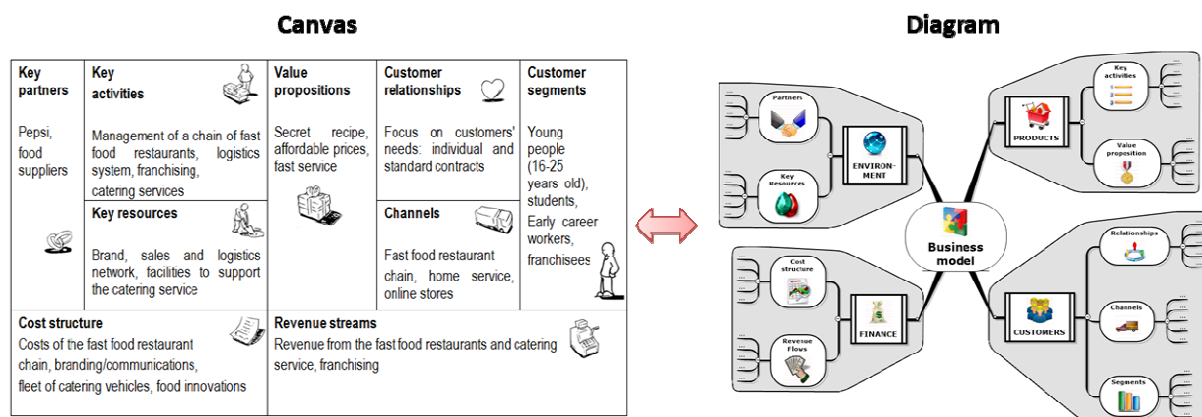


Figure 2. Business model representation in canvas and diagram

The 2 forms of representation (table vs map) were evaluated by 16 managers from executive education cohort. The obtained results allow concluding that as a way of representing a business model, a mind map is more effective than a business model canvas in terms of several important factors: easiness to use, speed of perception, clarity and understandability, aesthetic pleasure, opportunity to use in operational activities. The highest difference was seen in the criteria “opportunity to use of operational activities”. This major difference can be explained by the fact that rich functionality of contemporary mind mapping software (e.g., MindManager, iMindmap, iThoughts) facilitates using of the visualized business models for strategic management.

The second example is based on [Kudryavtsev et al, 2014, a, b]. The papers suggest the model-oriented method for business architecture alignment, which uses proven matrix-based Quality Function Deployment (QFD) methodology for analysis, decision making and communication. The central element of this method is the matrix, which is called “The House of Quality” [Hauser, Clausing, 1988]. This matrix method can augment famous visual method of strategic planning – Kaplan and Norton’s Strategy Maps [Kaplan, Norton, 2004], see Fig. 3. Strategy map in this example is better for capturing “big picture” and strategy communication, while matrix-based representation is better for detailed analysis of relationships between objectives (elements of strategy map).

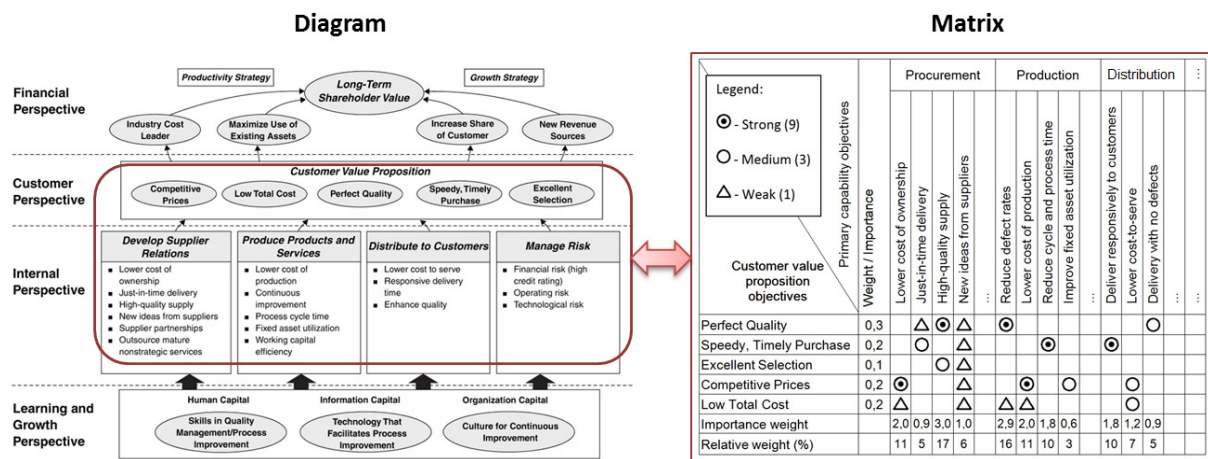


Figure 3: Strategy map and corresponding matrix with relationships

### Multi-representation of organizational knowledge using ontologies

Organizational knowledge for strategic management can be structured and processed by specialized enterprise architecture management tools. Enterprise architecture (EA) is an approach to provide insight and overview in order to manage the complexity of an organization and to aid strategic decision making [Op't Land et al, 2009]. EA is based on enterprise modeling and implies documentation of enterprise strategies, business capabilities, business processes, organizational structures, and information technologies, and especially their interaction and dependencies. From representation perspective EA models include catalogs, matrices and diagrams [TOGAF, 2011]. Originally, EA was developed as a tool for information systems management [Kappelman et al, 2008]. During the previous decades the concept has evolved more towards an instrument for business IT alignment [Simon et al, 2013]. EA has included business goals, value chain, business capabilities etc. as elements since it was first introduced by Zachman [Zachman, 1987] in the late 1980s. Now EA is more and more attached to enterprise transformation [Labusch, Winter, 2013] and strategic management [Aldea et al, 2013; Simon et al, 2013]. Simon et al. [Simon et al, 2013] show that EA could support the strategic planning process in several phases. According to them, EA would be most valuable in the strategy formulation and implementation phases, when assessing the readiness of the organization for transformation and deciding on how to execute the chosen strategy. Furthermore, they show that EA is least valuable in the strategy review phase. This is because the final performance can have been impacted by a variety of soft factors such as the employee resistance to change, which cannot be measured with the aid of EA. Enterprise architecting is supported by corresponding tools [Bittler, 2012]. Enterprise architecture management tools not only capture relevant information, but also process this information, e.g. using reports, visualizations or applying analytical methods.

As we've mentioned diagrams and matrices/tables have their own advantages and disadvantages. It is important to find cognitive fit between task, type of information and representation format. The same information should be represented in different formats depending on the task and context. The ontology-based approach for enterprise architecture management can be used for this purpose. Ontology is a formal, explicit specification of a shared conceptualization [Studer et al, 1998]. A 'conceptualization' refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. 'Explicit' means that the type of concepts used, and the constraints on their use are explicitly defined. For example, in medical domains, the concepts are diseases and symptoms, the relations between them are causal and a constraint is that a disease cannot cause itself. 'Formal' refers to the fact that the ontology should be machine readable, which excludes natural language. 'Shared' reflects the notion that ontology captures consensual knowledge, which is not private to some individual, but accepted by a group. Ontologies and corresponding semantic technologies are actively used for information integration, knowledge management, e-commerce, education and semantic web [Domingue et al, 2011; Gavrilova, Laird, 2005; Gomez-Perez et al, 2003; Gorovoy, Gavrilova, 2007]. Ontologies are also used for enterprise modeling, but these applications are mostly geared towards business process modeling/management and data integration. Ontologies for business architecture modeling, visualization and reporting are not yet applied.

It is suggested to use ontology as a metamodel for enterprise models. A populated enterprise ontology is equal to an enterprise model. All the necessary stakeholders' concerns are satisfied using ontology-based views. These views can be either document-oriented (text, table) or visual (diagram). The contents and the form of these views are defined using specifications (or viewpoints). Figure 4 represents the transition from the collection of independent diagrams and tables to the mapping between the diagrams, matrices and the enterprise ontology. This mapping provides the translation of ontology-based enterprise model into the partial views. Similar ideas and methods are currently being discussed in the “Semantic Cartography” community: <https://www.linkedin.com/groups/8101187> and are organized by Bernard Chabot on his website <https://www.topincs.com/SemanticCartography/1345>.

---

### **Examples of ontology-based tools for multi-representation of organizational knowledge**

---

The aforementioned idea is implemented in the following two technologies.

The ORG-Master modeling approach has originally been conceived in the course of the development of the business engineering toolkit in 1998 [Kudryavtsev et al, 2006; Grigoriev, Kudryavtsev, 2011; Grigoriev, Kudryavtsev, 2013]. Fig. 5 represents the suggested technology.

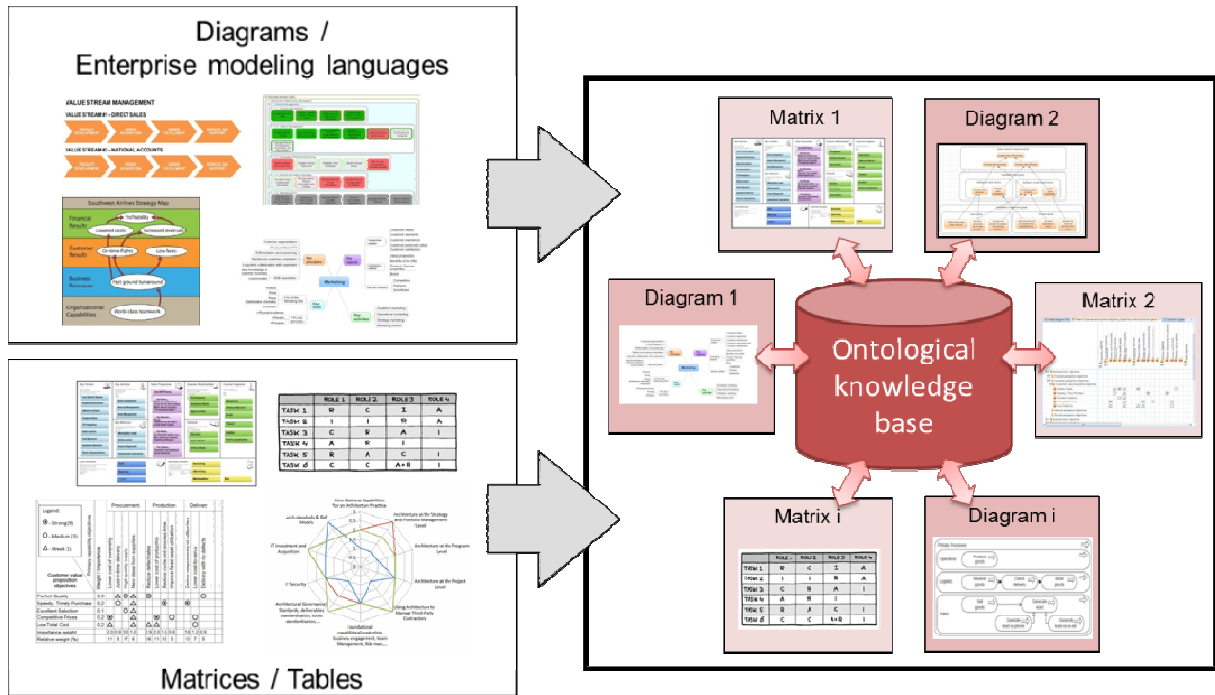


Figure 4. Multi-representation of organizational knowledge using ontologies

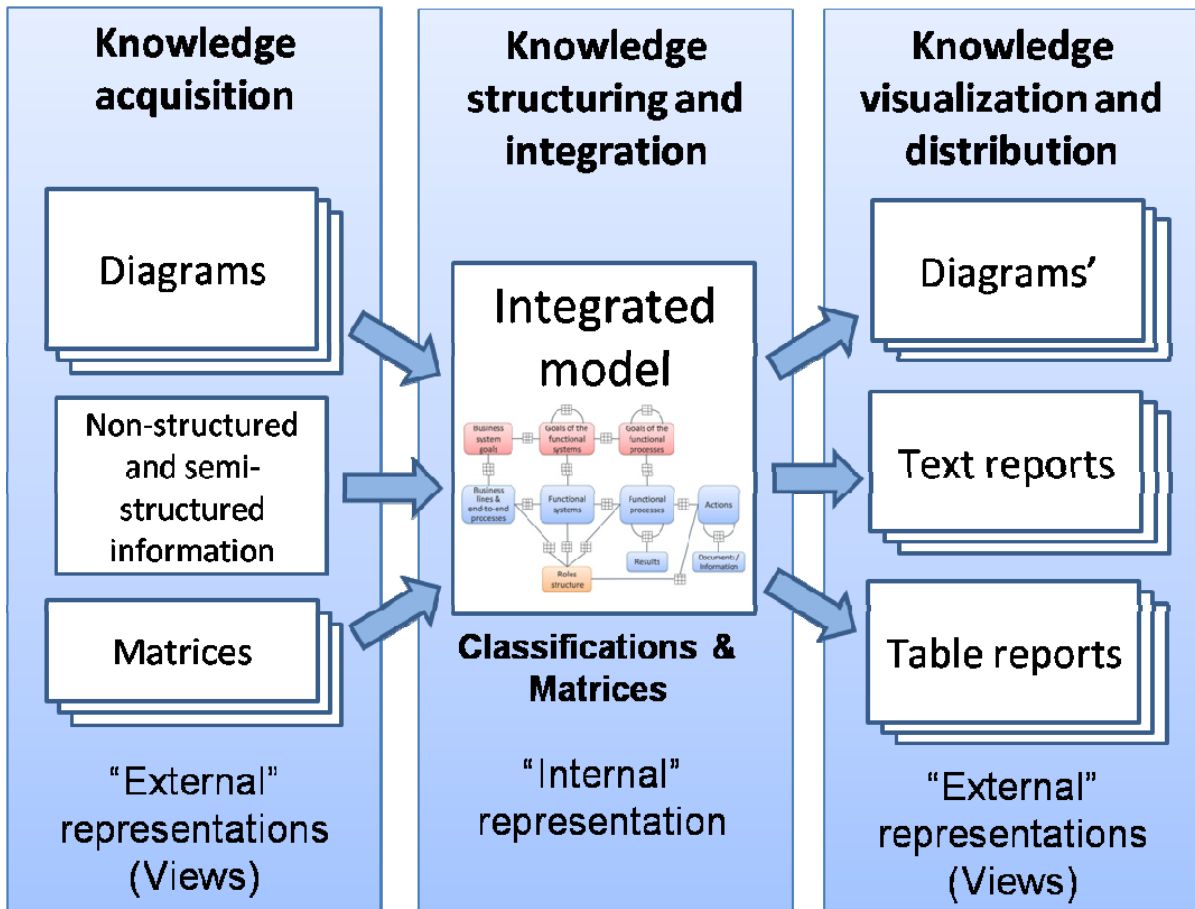


Figure 5. ORG-Master modeling approach

The non-diagrammatic method plays the central role in this technology. The modeling process starts from knowledge acquisition. Enterprise modelers collect information about an enterprise from various sources (people's memory, documents etc.), then organize it using diagrams, classifications and matrices. Diagrams can be created in ORG-Master graphical editors, which are based on Microsoft Visio. Typically process flow diagrams, concept maps, strategy maps and organizational chart are in use during business architecture engineering. These artifacts can be discussed and agreed upon with managers (managers can also make models by themselves). This step is standard for EAM tools. Then ORG-Master integrates all the acquired knowledge using classifications (hierarchical lists) and matrices – so called “internal” representation. These integration and complex structuring is typically done by highly qualified modelers. This step helps to provide holistic big-picture and consistency in large-scale EA models. However the resultant integrated model is inappropriate for final users and enterprise stakeholders, so ORG-Master provides capabilities to specify and generate partial views from this model (diagrams, text and table reports), which will suit various concerns of various stakeholders. Consistency of the “internal” model and of the “external” views is achieved through automatic model transformations, which are based on a shared unified metamodel (enterprise ontology) and mappings (between different notations and shared metamodel).

The second technology – Essential project – is suggested in [Mayall, Carter, 2015]. The Essential Project [2016], a ten-year development program that has produced an open source enterprise architecture support toolkit with a comprehensive metamodel. In common with other enterprise architecture management suites, Essential enables users to define and describe their enterprise in terms of its current and future states. The Essential Project is the collective name for a set of open source, enterprise architecture support tools that have been developed for use in conjunction with a variety of Enterprise Architecture approaches and frameworks.

More specifically, the components that currently comprise the Essential Project are:

- The Essential Meta-Model, a framework-independent set of semantic definitions for knowledge related to the building blocks and relationships of an enterprise.
- The Essential Architecture Manager, a knowledge repository and reporting tool for capturing and then querying information based on the Essential Meta-Model.

The Essential Architecture Manager is a toolset that is focused solely on supporting enterprise architecture practices, applied in the context of a variety of business and IT management processes (e.g. strategy management, IT governance, solution delivery, service delivery). The toolset offers all the required features of an enterprise architecture tool as defined by the Gartner Group. Fundamentally, these features can be grouped into two areas of functionality; functions that support users in the modeling of an enterprise, and functions that provide users with discrete views of this model in support

of reporting and analysis. This grouping is reflected in the underlying design of the toolset in that it comprises two main components, which separate the capture of information from the analysis (Fig. 6):

- Essential Modeller, providing support for capturing and maintaining the enterprise architecture model
- Essential Viewer, responsible for generating reports that allow users to view and analyse the enterprise architecture model. These reports can be in different forms: text, tables and diagrams.

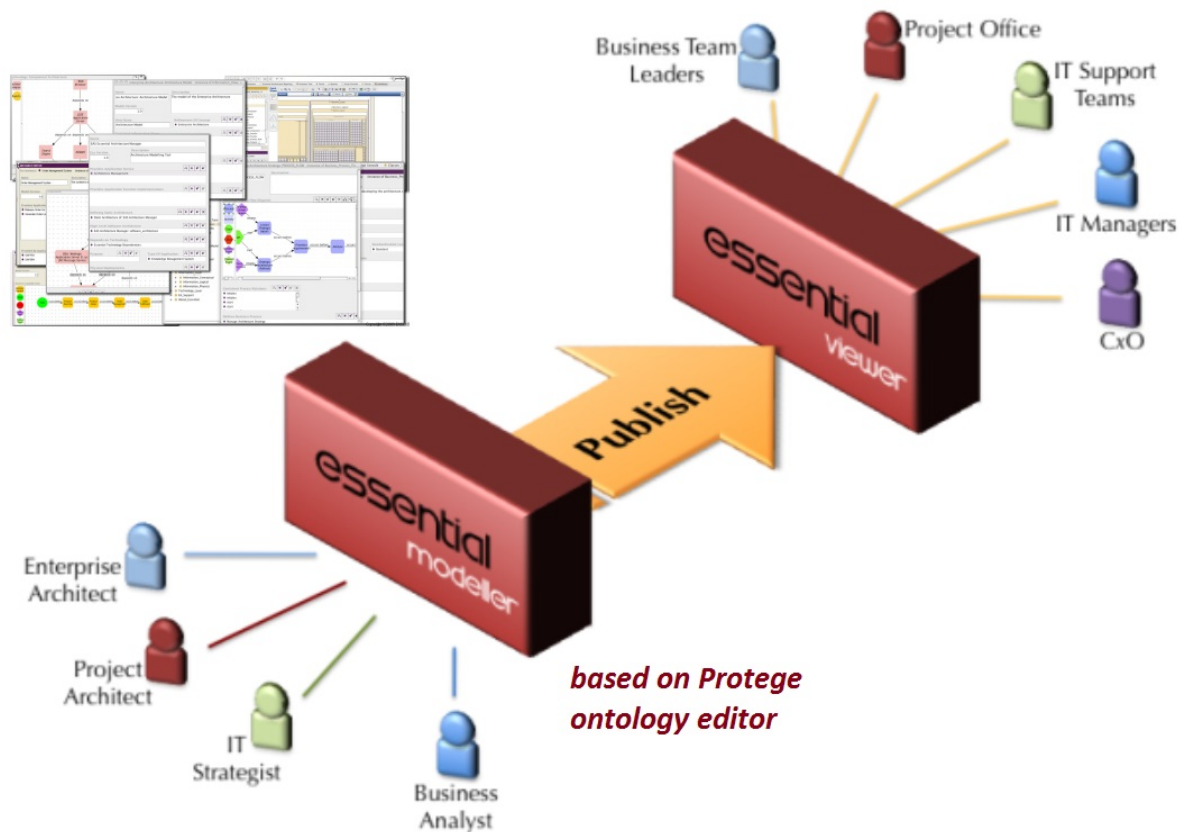


Figure 6. The Essential Architecture Manager [Essential project, 2016]

## Conclusion

The paper discusses methods and tools for organizational knowledge representation in the field of strategic management in the form of diagrams, matrices and ontologies. These three types of tools are used in strategic management for different purposes. Diagrams are better suited to solve problems associated with the managerial thinking, communications and coordination. Matrix tools can improve decision-making and elaborate analysis in strategic management. In order to provide multi-representation of organizational knowledge the third tool described in the paper could be used – ontologies. Ontologies are actively used for information integration, knowledge management and analytics. Two technologies are described in the paper as examples of ontology-based tools: ORG-Master modeling approach, which helps to collect information about an enterprise from various sources,

organize it using classifications and matrices, and generate partial views from them (including diagrams and matrices) suiting concerns of various stakeholders; and Essential project, which provides meta-model and software tool to develop ontology-based enterprise models and publish them in various forms.

---

### Acknowledgements

---

Research has been conducted with financial support from Russian Science Foundation grant (project No. 15-18-30048).

---

### Bibliography

---

- [Aldea et al, 2013] A. Aldea, M. E. Iacob, D. Quartel, H. Franken. Strategic planning and enterprise architecture. In: Enterprise Systems Conference (ES). 2013, pp. 1-8.
- [Bittler, 2012] S. Bittler. Magic Quadrant for Enterprise Architecture Tools, ID G00234030, Gartner Inc. 31 Oct. 2012, 28 p.
- [Buzan, 2003] T. Buzan. The Mind Map Book, BBC Active, London. 2003.
- [DeSanctis & Jarvenpaa, 1985] G. DeSanctis, S. Jarvenpaa. An investigation of the 'tables versus graphs' controversy in a learning environment. In L. Gallegos, R. Welke, & J. Wetherbe, (Eds.), Proceedings of the 6th International Conference on Information Systems, 1985. pp. 134-144.
- [Domingue et al, 2011] J. Domingue, D. Fensel, J. Hendler. (Eds.). Handbook of semantic web technologies. Springer Science & Business Media. 2011.
- [Eppler & Platts, 2009] M. Eppler, K. Platts. Visual strategizing: the systematic use of visualization in the strategic-planning process. In: Long Range Planning. 2009, 42(1), pp. 42-74.
- [Essential project, 2016] The Essential project official website. Available from: <http://www.enterprise-architecture.org/> [Accessed 23 March 2016].
- [Gavrilova, Alsufyev, Yanson, 2014] T. Gavrilova, A. Alsufyev, A.-S. Yanson. Modern Notation of Business Models: Visual Trend. In: Foresight-Russia. 2014, 8(2), pp. 56–70.
- [Gavrilova, Gulyakina, 2011] T. Gavrilova, N. Gulyakina. Visual Knowledge Processing Techniques: a Brief Review. In: Scientific and Technical Information Processing. 2011, 38 (6), pp. 403–408.
- [Gavrilova, Laird, 2005] T. Gavrilova, D. Laird. Practical Design of Business Enterprise Ontologies. In: Industrial Applications of Semantic Web. Eds. Bramer M. and Terzyan V. Springer. 2005, pp.61-81.
- [Ghoniem et al, 2005] M. Ghoniem, J. Fekete, P. Castagliola. On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. In: Information Visualization. 2005, 4(2), pp. 114–135.



- [Gomez-Perez et al, 2003] A. Gomez-Perez, O. Corcho, M. Fernandez-Lopez. *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. First Edition, Springer. 2003.
- [Gorovoy, Gavrilova, 2007] V. Gorovoy, T. Gavrilova. *Technology for Ontological Engineering Lifecycle Support*. In: *Information Theories and Applications*. 2007, 14 (1), pp. 19-25.
- [Grigoriev, Kudryavtsev, 2011] L. Grigoriev, D. Kudryavtsev. *Ontology-based business architecture engineering framework*. In: *Frontiers in Artificial Intelligence and Applications*. 2011, 231, pp. 233-252.
- [Grigoriev, Kudryavtsev, 2013] L. Grigoriev, D. Kudryavtsev. *ORG-Master: Combining Classifications, Matrices and Diagrams in the Enterprise Architecture Modeling Tool*. In: *Proceedings of the 4th Conference on Knowledge Engineering and Semantic Web, October 7-9, 2013*. *Communications in Computer and Information Science (CCIS) Series*, Springer. 2013, pp. 250-258.
- [Hauser, Clausing, 1988] J. Hauser, D. Clausing. *The House of Quality*. In: *Harvard Business Review*. 1988, 66 (May–June), pp 63–73.
- [Jarvenpaa & Dickson, 1988] S. Jarvenpaa, G. Dickson. *Graphics and managerial decision making: Research-based guidelines*. In: *Communications of the ACM*. 1988, 31(6), pp. 764-774.
- [Jarvenpaa, 1989] S. Jarvenpaa. *The effect of task demands and graphical format on information processing strategies*. In: *Management Science*. 1989, 35(3), pp. 285-303.
- [Kaplan, Norton, 2004] R. Kaplan, D. Norton. *Strategy Maps*. Harvard Business School Publishing, Boston, Massachusetts. 2004.
- [Kappelman et al, 2008] L. Kappelman, T. McGinnis, A. Pettite, B. Salmans, A. Sidorova. *Enterprise architecture: Charting the territory for academic research*. In: *AMCIS 2008 Proceedings*. 2008.
- [Keller et al, 2006] R. Keller, C. M. Eckert, P. J. Clarkson. *Matrices or node-link diagrams: which visual representation is better for visualising connectivity models?* In: *Information Visualization*. 2006, 5, pp. 62–76.
- [Kudryavtsev et al, 2006] D. Kudryavtsev, L. Grigoriev, V. Kislova, A. Zablotsky. *Using ORG-Master for knowledge based organizational change*. In: *Information Theories & Applications*. 2006, 13(2), pp. 131-139.
- [Kudryavtsev et al, 2014, a] D. Kudryavtsev, L. Grigoriev, S. Bobrikov. *Strategy-focused and value-oriented capabilities: methodology for linking capabilities with goals and measures*. In: *Proceedings of the 1st International Workshop on Capability-oriented Business Informatics (CoBI) as part of the 16th IEEE Conference on Business Informatics, Geneve, 14-17 July. 2014*, pp. 15-26.



- [Kudryavtsev et al, 2014, b] D. Kudryavtsev, L. Grigoriev, I. Koryshev. Applying Quality Function Deployment method for business architecture alignment. In: Proceedings of the 8th European Conference on IS Management and Evaluation (ECIME 2014), Ghent, Belgium. 11-12 September 2014, pp. 118-127.
- [Kudryavtsev, Gavrilova, 2016] D. Kudryavtsev, T. Gavrilova. From Anarchy to System: a Novel Classification of Visual Knowledge Codification Techniques. In: Knowledge and Process Management: The Journal of Corporate Transformation. 2016. In press.
- [Labusch, Winter, 2013] N. Labusch, R. Winter. Towards a Conceptualization of Architectural Support for Enterprise Transformation. In: ECIS. 2013, pp. 116.
- [Mayall, Carter, 2015] A. Mayall, J. Carter. The Essential Project: Harnessing Conceptual Structures to Expose Organizational Dynamics. In: International Journal of Conceptual Structures and Smart Applications (IJCSSA). 2015, 3(2), pp. 1-11.
- [Op't Land et al, 2009] M. Op't Land, E. Proper, M. Waage, J. Cloo, C. Steghuis. Enterprise Architecture Creating Value by Informed Governance, Berlin: Springer. 2009.
- [Osterwalder et al, 2005] A. Osterwalder, Y. Pigneur, C.L. Tucci. Clarifying Business Models: Origins, Present, and Future of the Concept. In: Communications of the Association for Information Systems (AIS). 2005, 16(1), pp. 1–25.
- [Osterwalder, Pigneur, 2010] A. Osterwalder, Y. Pigneur. Business model generation — A handbook for visionaires, game changers, and challengers, Wiley, New York. 2010.
- [Phaal et al, 2006] R. Phaal, C. Farrukh, D. Probert. Technology management tools: concept, development and application. In: Technovation. 2006, 26(3), pp. 336-344.
- [Simon et al, 2013] D. Simon, K. Fischbach, D. Schoder. An exploration of enterprise architecture research. In: Communications of the AIS. 2013, 32(1), pp.1–72.
- [Studer et al, 1998] R. Studer, R. Benjamins, D. Fensel. Knowledge Engineering: Principles and Methods. In: Data and Knowledge Engineering. 1998, 25(1-2), pp. 161-197.
- [TOGAF, 2011] TOGAF. Sample Catalogs, Matrices and Diagrams. 2011. Available from: <http://www.togaf.info/togafSlides91/TOGAF-V91-Sample-Catalogs-Matrices-Diagrams-v3.pdf> [Accessed 23 March 2016].
- [Vessey, 1991] I. Vessey. Cognitive fit: A theory-based analysis of the graphs versus tables literature. In: Decision Sciences. 1991, No 22, pp. 219–241.
- [Wickens, 1992] C. Wickens. The proximity compatibility principle: Its psychological foundation and its relevance to display design. Technical Report ARL-92/NASA-92-3. Savoy, Illinois: Aviation Research Laboratory, Institute of Aviation, University of Illinois at Urbana-Champaign, 61874. 1992.

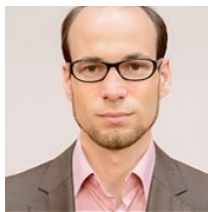
[Wickens, Merwin, & Lin, 1994] C. Wickens, D. Merwin, E. Lin. Implications of graphics enhancements for the visualization of scientific data: Dimensional integrality, stereopsis, motion, and mesh. In: Human Factors. 1994, 6(1), pp. 44-61.

[Zachman, 1987] J.A. Zachman. A framework for information systems architecture. In: IBM Systems Journal. 1987, 26(3), pp. 276–292.

---

### Authors' Information

---



**Dmitry Kudryavtsev** – PhD, Associate Professor, Information Technologies in Management Department, Graduate School of Management (GSOM) in St. Petersburg University; e-mail: [d.v.kudryavtsev@gsom.pu.ru](mailto:d.v.kudryavtsev@gsom.pu.ru).

*Major Fields of Scientific Research: knowledge structuring and representation, enterprise architecting, sensemaking and decision making.*



**Anna Menshikova** – PhD, Head of Research Projects Department at Graduate School of Management (GSOM) St. Petersburg University, Member of Research Team of INNOVARRA Project (Innovations in Organizational Knowledge Management: Typology, Methodology and Recommendations); e-mail: [menshikova@gsom.pu.ru](mailto:menshikova@gsom.pu.ru). Research interests: Discourse Analysis, Text/Image Relations, Cultural Studies, Management Education, Knowledge Management and Knowledge Sharing.



**Tatiana Gavrilova** – Professor, Head of Information Technologies in Management Department at Graduate School of Management (GSOM) in St. Petersburg University; e-mail: [gavrilova@gsom.pu.ru](mailto:gavrilova@gsom.pu.ru). Research interests: knowledge codification and structuring using the cognitive approach.

## METHODS AND ALGORITHMS OF TIME SERIES PROCESSING IN INTELLIGENT SYSTEMS

**Sergey G. Antipov , Marina V. Fomina, Vadim V.Vagin,  
Alexandr P. Ereemeev, Vasili A. Ganishev**

**Abstract:** *Time series processing in intelligent is a complex cross–disciplinary problem that is posed in many research areas. In this paper two subproblems are considered: anomaly detection in time series and time series clustering on an example of speaker clustering.*

**Keywords:** *Time Series Processing, Anomaly Detection, Speaker Clustering, Neural Networks.*

---

### Introduction

---

An Intelligent System (IS) is viewed as a computer system to solve problems that cannot be solved by human in real time, or a solution requires automated support. The solution should give results comparable to the decisions taken by a person who is a specialist in a certain domain. The most important class of problems whose solution requires the intelligent support is a complex technical object management. The main feature of such objects is that they are dynamic, have ability for developing, their state may change over time, so one needs to develop methods and algorithms that take into account a time factor. One of the basic tasks arising when processing temporal dependences is the task of a clustering and classification. The review of methods of such tasks solution will be given below. It is offered to consider a clustering problem on the example of the task of speaker recognition on a voice. The problem of classification is solved on the example of anomalies search in sets of time series.

The paper is structured as follows. In section 2, the concept of time series and problems of their processing are given. The most important problems arising in the case of time series analysis are considered. The classes of methods for their decision are numerated. Section 3 contains the review of the main clustering methods. In section 4, the clustering problem to audio signal processing is viewed. The offered method of a task solution and the practical implementation are described. Section 5 contains the description of the method and anomaly search algorithms in collections of time series. In subsection 5.1, setting up the anomaly search task is given. In subsection 5.2, the normalization of time series is presented. Subsection 5.3 represents search algorithms of exceptions in collections of time series for the cases of one and several classes. In subsection 5.4, results of computer modeling are given. In Section 6, there is given the temporal data model on the basis of which the temporal data

model for a subsystem of time series processing in intelligent systems of real time can be described. Conclusions are presented in section 7.

---

### The Problems of Processing Time Series

---

As for the analysis of complex technical objects behaviour requires the consideration of the time factor, there is a need to work with data that explicitly (or implicitly) contains time. In this regard, one has to deal with the problem of temporal data mining [Roddick and Spiliopoulou, 1999, Weiqiang et al., 2002, Antunes and Oliveira, 2001]. The most common case of this analysis is time series mining [Perfilieva et al., 2013]. Time series are used in many different areas (technics, economics, medicine, banking, etc.) and describe different processes that occur over time.

The problem of time series mining is important for solving the following tasks of process analysis.

1. A process state prediction depending on the qualitative evaluation of a current or previous state.
2. Abnormal event detection.
3. Time series trends or other change identification.

The following classes of methods are used to solve these problems: Associations, Sequence, Classification, Clustering methods.

In combination with other methods of data mining, **prediction** or **forecasting** involves trend analysis, classification and model matching. The basis for all kinds of forecasting is the historical information stored in the database in the form of time series. If one can build or find patterns, that adequately reflect the object behaviour, it is likely that they can be used to predict the behaviour of a system in the future.

The problems of detecting trends, their qualitative assessment and forecast based on analysis of time series are of particular relevance due to the continuous growth of real-time data from the specific and complex technical objects, for example, sensors whose values change over time.

Consider the case where the object's behaviour is evaluated on the basis of particular parameter values observations.

In general, the time series  $TS$  is an ordered sequence of values  $TS = \langle ts_1, ts_2, ts_i, \dots, ts_m \rangle$  describing the flow of a long process, where the index  $i$  corresponds to a time mark.  $ts_i$  values can be sensor indications, product prices, exchange rates and so on.

---

### Clustering Methods

---

The time series clustering problem belongs to a class of pattern recognition problems [Vagin et al., 2008]. There are the following types of methods: hierarchical; methods based on statistical analysis and machine learning. The most commonly used methods are:

- agglomerative hierarchical clustering with increasing mixture of normal (Gaussian) distributions;

- Hidden Markov models;
- Kohonen networks;
- Histogram models;
- incremental self-organizing neural networks.

### ***Agglomerative hierarchical clustering with increasing mixture of normal (Gaussian) distribution***

A cluster corresponding to the time series can be described as a probability distribution. However, very often a cluster has a rather complicated shape, that can not be described by a single probability distribution. In such cases, a mixture of probability distributions is used for the description of a cluster.

For the exact cluster description, it is necessary to learn a mixture of distributions. For this very often the **EM-algorithm (Expectation-Maximization)** is used [Zhu et al., 2005]. We introduce an auxiliary vector of hidden variables  $G$  that can be calculated for the known value of a parameter vector  $\Theta = \{\theta_1, \dots, \theta_s\}$ . In its turn, a vector  $G$  helps to restore a vector  $\Theta$ . The algorithm consists of two iterative repetition of steps [Roweis, 1998].

1. In the first step (**E-step**) the expected value of hidden parameters of a vector  $G$  is calculated for the current approximation of the vector  $\Theta$ .
2. In the second step (**M-step**) the problem of maximizing the likelihood for a mixture of distributions is being solved, and a new vector  $\Theta$  approximation is computed.

In the case when all mixture components have normal (Gaussian) probability density, one can represent the solution analytically. That is why mixtures of normal distributions (GMM - **G**aussian **M**ixture **M**odel) are used commonly in practice.

*Agglomerative hierarchical clustering (AHC)* is one of the most popular clustering methods, since it is simple to implement, but provides the accuracy sufficient for many applications. The basic idea is that at initialization, each time series are represented by separate clusters. Then, some clusters close one to another are merged together. This process is as long as certain criterion indicates that further merge does not lead to better results.

As a measure of the distance between clusters often distance-based Bayesian information criterion (BIC) is used.

However, the EM-algorithm can not be used at the early stages of the AHC, since most of initial clusters does not generally contain enough information for learning multicomponent GMM. The result is an overfitted model, which is correct only for the particular case.

In view of these shortcomings the method based on the incremental mixture of normal distributions (incremental GMM) was proposed in [Han and Narayanan, 2008]. The method is characterized by the following:

- a cluster, that is a union of two adjacent clusters, is represented by the distribution whose probability density is calculated as a weighted sum of merged clusters;
- the model is recursively updated after each combination of clusters using the hypothesis  $H_2$ , where the distance between clusters reaches a predetermined value.

Owing to the incremental GMM usage, researchers succeeded in increasing the quality of clustering at 4.47% [Han and Narayanan, 2008].

### **Hidden Markov models**

Hidden Markov model (HMM) is a Markov process model, where the system initiating a process is in any state from  $C = \{c_1, c_2, \dots, c_M\}$ , but it is not known which one exactly. However, each of the state  $c_i$  with the likelihood  $b_{ix_j}$  produces an observed event  $x_j$ .

Formally HMM is defined as  $HMM = \{C, X, \Pi, A, B\}$  [Rabiner, 1989, Abdallah et al., 2012], where:

- $C = \{c_1, c_2, \dots, c_M\}$  is a system states set (clusters);
- $X = \{x_1, x_2, \dots, x_N\}$  is a cluster centroids set;
- $\Pi = \{\pi_1, \pi_2, \dots, \pi_M\}$  is a list of initial (a priori) probabilities of membership to each cluster;
- $A = \{a_{ij}\}$  - the matrix of transitions between events: to which a cluster belongs the next feature vector;
- $B = \{b_{ix_c}\}$  is the matrix defined a communication probability of a feature  $i$ -s vector with a centroid  $x_c$  into a cluster  $c_i$ .

Each state consists of a set of substates. The probability density of each state is modelled by a GMM.

At absent of a priori knowledge about the number of time series types, the quantity of clusters is selected more than the maximum possible (over-clustering). In this approach, multiple clusters would correspond to one and the same type of time-series, that leads to the necessity of merging in the later stages of clustering. Let us describe the clustering algorithm.

In the first step of this algorithm, one needs to initialize HMM parameters. At this stage, one can use the classical  $k$ -means algorithm. For each cluster, GMM parameters are computed.

Then it is necessary to learn HMM using the **EM-algorithm** of clusterization.

On the second step of the algorithm it is necessary to converge several clusters belonging to the same type of time series to one. It is quite difficult to determine the optimal number of clusters analytically, so in practice one chooses  $k = K$ , where  $K$  is a value greater than the maximum possible number of clusters. In the learning process, the current cluster quantity  $k$  is reduced to  $k^*$ , the optimal number of clusters. Assuming that the correct cluster merge increases the value of the objective function and incorrect merge decreases it, one can use the BIC.

In [Ajmera and Wooters, 2003] it is shown that the BIC for this method can be optimized. Assumption is entered that an amount of parameters of every GMM must be a constant. This assumption greatly increases the speed of HMM learning.

### **Kohonen networks**

In recent years, the usage of neural networks for time series clustering is of growing interest [Ning et al., 2006], [Mori and Nakagawa, 2001]. In particular, Kohonen network is frequently used because of the high rate of clustering.

This method is based on the projection from a multidimensional space in a two-dimensional one with some predefined structure. For Kohonen network learning, the Linde-Buzo-Gray (LBG) algorithm is often used [Linde et al., 1980]. This algorithm helps to divide L feature vectors corresponding to the same time series type to M clusters. A centroids (cluster centres) set is called “codebook”, it is unique for each time series type.

Several centroids in a codebook will match each time series type. In clustering step, a feature vectors are included to a cluster having the nearest centroid to this vector in its codebook [Kumar and Rao, 2011].

### **Incremental self-organizing neural networks**

To remove the previous model limitations, self-organizing incremental neural networks (SOINN) have been developed [Furao and Hasewaga, 2006]. These neural networks are used for data clustering without a priori knowledge of its topology. Also, the model supports the learning without the ultimate goal for the whole period of the network operation (lifetime learning), it allows not to limit the maximum number of clusters. SOINN is the neural network with two layers. The first layer is used to determine the topological structure of the clusters, and the second is used to define the clusters number. First, the first layer is learned and then using the output of the first layer as an input, a second layer is trained. Researchers need to identify themselves the time to start learning the second layer, as well as to calculate the threshold T for each layer of the network. In the absence of a priori data structure knowledge, the threshold for the first layer is often selected adaptively.

The main idea of the algorithm is to build a probabilistic model of an input data. On the assumption that clusters form an area of high probability density, it is necessary to construct a graph that most accurately describes these areas and their relative positions in space. Its nodes lie in the areas of local likelihood maximum, and its edges join nodes belonging to the same cluster.

Large number of parameters that SOINN has and uncertainty in choosing a time to start the second layer learning makes this type of networks difficult to use in practice. Also, if there is any change in the first layer then the second layer must be completely retrained. It makes online learning or lifetime learning impossible.

To solve the above problems enhanced self-organizing incremental neural network (ESOINN) were developed [Furao *et al.*, 2007]. The main difference of this approach is the use of single-layer neural network, which reduces the number of configurable settings.

---

### Clustering Problem in Task Of Speaker Clustering

---

#### **Problem Statement**

Let us consider the problem of time series clustering on the example of audio signal processing.

Modern recording tools allow to introduce a sound signal into a time series, showing the change in sound intensity over time. However, this view is difficult to analyze because it contains a large amount of information noise. A signal spectrum is more informative for analysis than the signal itself. For calculation of the spectrum the Fast Fourier Transform algorithm is often used. It is easy to implement and has a complexity  $O(N \log_2 N)$  less than the classical discrete Fourier transform algorithm  $O(N^2)$  [Cooley and Tukey, 1965].

Speaker clustering is a separation of voice recordings to some classes so that each class has only a voice of one user. Each recording contains a voice of a single person. This process is often an integral part of speaker recognition and speech recognition problems.

During evolution sounds in the lower frequency band contained the more useful information than ones in the higher frequency band. Mel-frequency cepstral coefficients (MFCC) were developed according to these characteristics of human hearing [Vyas and Kumari, 2013]. With these coefficients the information obtained from the low-frequency range is more carefully analysed, and the effect of high-frequency components typically containing extraneous noise is reduced.

Whole voice recording is divided into small intervals with duration  $\sim 10\text{-}30\text{ms}$  (signal quasi-stationary time) called frames. For each frame separately a set of mel-frequency cepstral coefficients is calculated.

The algorithm for computing mel-frequency cepstral coefficients can be described as follows [Molau *et al.*, 2001]:

- a) splitting a signal into frames;
- b) an application of the weighting function (window) to each frame;
- c) a use of the Fourier transform;
- d) a use of mel-frequency filter;
- e) a cepstrum calculation.

But MFCC change also contains unique information about the user's identity. In this work an extension of the previous acoustic vector by taking into account the dynamics of MFCC  $\delta_i$  is used, which is expressed by the difference in a mel-frequency cepstral coefficients of the frame  $i$  and the previous one:



$$\delta_i(\tilde{C}_k) = \tilde{C}_k[i-1] - \tilde{C}_k[i]$$

Here each  $\tilde{C}_k$  is a set of received mel-frequency cepstral coefficients,  $k = 1, \dots, K$ .  $K$  is quantity of mel - coefficients, the value of  $K$  is often chosen in a range from 12 to 24.

In this approach the first frame can not be used for clustering, since a change of its MFCC will be zero.  $L$  (the number of elements of the acoustic vector  $x$ ) doubles:

$$L = |x| = \left| \left[ \tilde{C}_1, \dots, \tilde{C}_k, \delta(\tilde{C}_1), \dots, \delta(\tilde{C}_k) \right] \right| = 2 \cdot K$$

A set of acoustic vectors will to be used for speaker clustering with ESOINN.

### **Practical implementation**

For MFCC calculation freeware package Praat was used, it was developed at the University of Amsterdam [Boersma and van Heuven, 2004]. This tool implements many features required for speaker and speech recognition and has a simple interface and constant developers' support.

After building MFCC sets advanced acoustic vectors containing the information about the dynamic change of MFCC are computed based on them.

CMU Sphinx project [CMU Sphinx] provides three ready-made base of high quality voice recordings: CMU Arctic; CMU Chaplain; CMU Microphone Array Database.

The first two voice recording bases were used in the experiments for this paper.

CMU Arctic base is established in Language Technology Institute at Carnegie Mellon University [CMU Arctic] and is divided into 4 parts, each contains only one user speech in English. Records are presented in the format of Wave with 16 kHz sampling frequency and EGG, each of them is phonetically balanced.

Base CMU Chaplain was developed at Carnegie Mellon University together with the Lockheed Martin System Integration as part of a hardware and software system for automatic translation (Audio Voice Translation Guide System, also known as Tongues) [CMU Chaplain]. Recordings include a dialogue between two chaplains in the English, that lasts 4 hours 15 minutes.

For each frame, a set of 13 MFCC was calculated. Based on that the acoustic vector with 26 components was built.

To assess the quality of clustering a concept of precision  $A$  will be used (proportion of recordings correctly associated with the users), and the error  $E$  (the proportion of recordings incorrectly associated with the user), defined as follows:

$$V = V_T + V_F, \quad A = \frac{V_T}{V}, \quad E = \frac{V_F}{V},$$

where  $V_T$  is the number of recordings correctly associated with the user, and  $V_F$  is the number of recordings incorrectly associated with the user.

The accuracy of speaker clustering method for each database CMU Arctic and CMU Chaplain can be represented by the following table:

**Table 1.** Experimental Results

Number of recordings	CMU Arctic	CMU Chaplain
100	84%	88%
200	86,5%	89,5%

It is worth noting that in the case of 100 recordings from CMU Arctic the system has identified five different users instead of four, but it is not mentioned in the calculation used for the accuracy of clustering users voice formula.

---

### Anomaly Detection

---

The anomaly detection problem [Chandola *et al.*, 2009] is set up as the task of searching for patterns in data sets that do not satisfy some typical behaviors. The ability to find abnormalities in a data set is important in a variety of subject areas: in the complex technical system analysis (e.g. satellite telemetry), in network traffic analysis, in medicine (analysis of MRT images) in the banking industry (fraud detection) and etc.

The anomaly, or "blowout" is defined as an element that stands out from the data set which it belongs to and differs significantly from the other elements of the sample. Informally, the problem of anomaly detection in time series sets is formulated as follows. There is a collection of time series describing some processes. This collection is used to describe normal processes. It is required to construct a model on the basis of the available data, that is a generalized the description of normal processes and allows to distinguish between normal and abnormal processes.

The task is complicated by the fact that a set of input data is limited and does not contain any examples of abnormal processes. It does not specify a criteria by which it would be possible to distinguish the <normal> and <abnormal> time series. In this regard, it is difficult to accurately assess the quality of the algorithm (the percentage of correctly detected anomalies, the number of false positives and the number of missed abnormalities). In addition, many algorithms are working well for some data sets will not fit well for other subject areas. It may also vary a criteria for determination the <correct> time series.

Let there be a set of objects, where each object has a time series:

$TSSstudy = \langle TSstudy_1, TSstudy_2, \dots, TSstudy_m \rangle$  is a learning set. Each the time series in the learning set is an example of <normal> process flow. Based on the analysis of time series of  $TSSstudy$  one needs to build a model to refer the testing set of time series  $TSTEST = \langle TStest_1, TStest_2, \dots, TStest_m \rangle$  to <normal> or <abnormal> by some criterion.

This problem should be divided into two cases: the first case, when learning set contains examples of a single class; the second case, when the learning set contains examples of several classes. In the first case the fact of member of these objects to the class of the training set is important. In the second case one needs to further define an object belongs to a particular class.

Here are the main methods used to solve the problem of classification.

Classification is used to learn the model for the data assigned to different classes (learning stage), and to refer instances to one of the existing classes using the resulting model (test stage). Anomaly detection methods are based on the classification, it is assumed that if a classifier can be learned in an existing feature space, it can separate the normal and abnormal objects.

The advantage of anomaly detection methods based on classification includes the ability to use a lot of techniques and algorithms developed in the machine learning field, especially in the case where the learning set contains examples of several classes. Further test stage is fast compared to other classes of methods, as used originally constructed model (classifier).

### **Problem statement**

Let there be a set of objects, where each object is a time series:

$TSSstudy = \langle TSstudy_1, TSstudy_2, \dots, TSstudy_m \rangle$  is a learning sample. Each of time series in the learning set is an example of a "normal" process flow. Based on time series mining from a set  $TSSstudy$  one needs to build a model to refer time series from the test sample  $TSTEST = \langle TStest_1, TStest_2, \dots, TStest_t \rangle$  to "normal" or "anomalies" on the basis of some criterion.

Let us consider this problem with a simple example. Let  $TSSstudy$  learning set consists six time series (Fig. 1).

Test sample  $TSTEST$  consists of three time series (Fig. 2).

Based on the above problem statement it is clear that the time series (1), (2) and (6) of a study set are highly similar to each other, and therefore are members of the same class, let it be Class 1. Time series (3), (4) and (5) are also similar, but belong to another class, let it be Class 2. The test set (Fig. 2) shows that the time series (1) is likely to be a member of Class 1, a time series (2) is a member of Class 2. The third time series is significantly different from the previous two, and apparently "not similar" to any in the

learning set. This suggests that the process by which time series (3) from the test sample was received is different from processes, by which time series from the learning set were obtained. On the contrary, the time series (1) and (2) from the test set (Fig. 2) are not anomalies, since their shapes are very "similar" to the individual time series in the learning set.

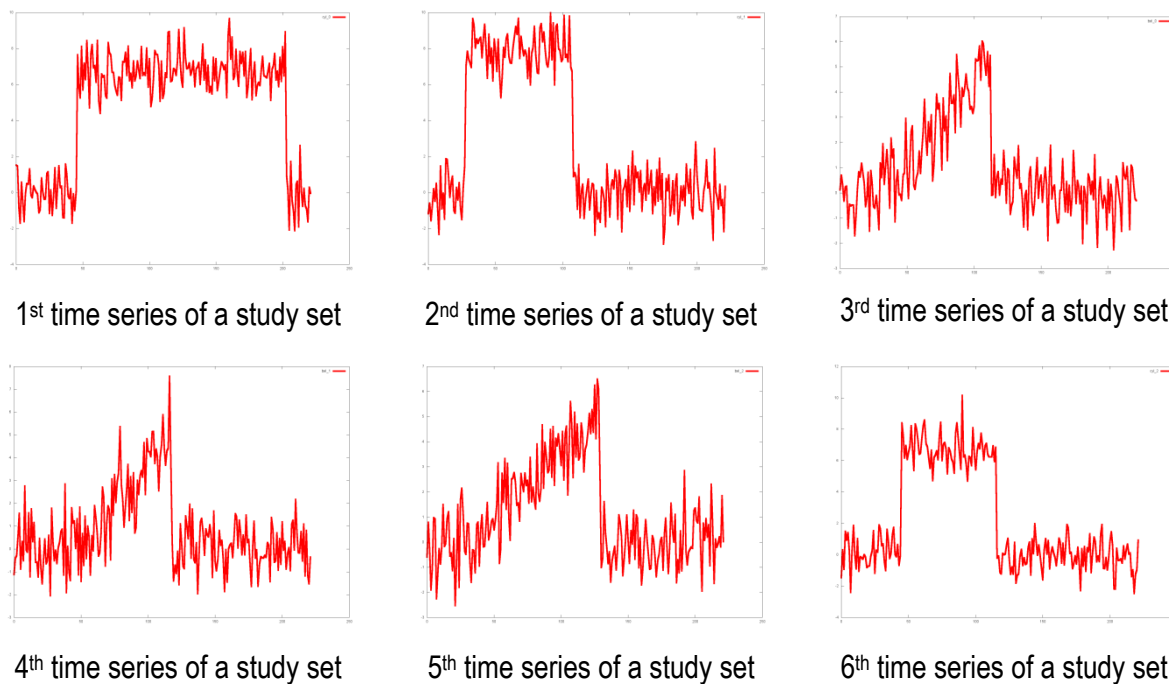


Figure 1. An example of a study set

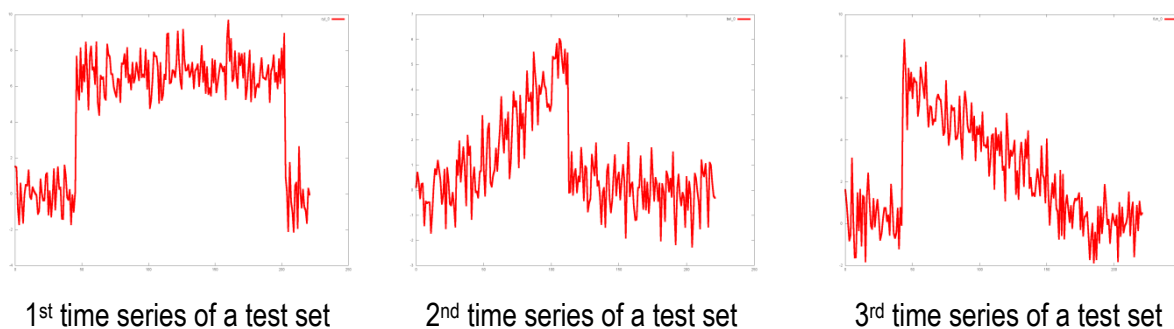


Figure 2. An example of a test set

**Methods of time series representation**

To create an algorithm summarizing an information provided by time series it is required, of course, to develop methods of time series pre-conversion. It is required to cause time series, that represent data from different areas, in different units to some forms convenient for further analysis. To work with time series, it is proposed to use two methods of representation a normalized and a symbolic representation. Normalization is bringing time series to a form that its mean value would be equal to zero and standard deviation would be one; this transformation is a necessary process of the data pre-processing [Lin et al., 2003]. Examples of current and normalized time series are shown in Table 2 (lines 1 and 2)

**Table 2.** Current and normalized time series representation

Time $t$	1	2	3	4	5	6	7	8	9
Absolute values	512	1448	88	1448	1448	1448	1448	1024	512
Normalized values	-1.0415	0.7478	-1.852	0.7478	0.7478	0.7478	0.7478	-0.0627	-0.0415
Symbolic representation of normalized values	C	P	A	C	C	C	C	J	C

Symbolic time series representation can be obtained from normalized one using the algorithm Symbolic Aggregate approxIimation [Lin et al., 2003]. To perform this conversion the alphabet  $\mathbf{A}$  - finite set of characters is introduced:  $\mathbf{A} = \{a_1, a_2, \dots, a_{|\mathbf{A}|-1}\}$ .

An example of symbolic representation for the time series is given in Table 2 (line 3). Alphabet  $\mathbf{A}$  was examined consist of 20 characters,  $\mathbf{A} = \{A, B, C, \dots, T\}$ .

**Anomaly Detection Algorithm in time series sets**

This paper proposes a method for the anomaly detection in the sets of time series, that is a modification of a method based on an accurate exceptions description [Arning and Agrawal, 1996]. The original formulation of this problem is given in [Arning and Agrawal, 1996]:

for a given set of objects  $I$  one needs to get an exclusion-set  $I_x$ . To do this, set for set  $I$  are introduced:

The function of dissimilarity  $D(I_j, I_j) \in I$  defined on  $P(I)$ , the set of all subsets of  $I$  and receiving positive real value;

The cardinality function  $C(I_j): I_j \subseteq I$ , defined on  $P(I)$  and receiving positive real value such that for any  $I_1 \subset I, I_2 \subset I$  performed  $I_1 \subset I_2 \Rightarrow C(I_1) < C(I_2)$ ;

The smoothing factor  $SF(I_j) = C(I \setminus I_j) \cdot (D(I) - D(I \setminus I_j))$ , which is calculated for each  $I_j \subseteq I$ . Then  $I_x \subset I$  will be considered an exclusion-set for  $I$  with respect to  $D$ , and  $C$ , if its smoothing factor  $SF(I_x)$  is maximal [Arning and Agrawal, 1996].

The algorithm TS-ADEEP that is based on this method was adapted for anomaly detection problem in sets of time series. As set  $I$  sets  $TSStudy \cup \{ts\_test_j\}$  for each  $ts\_test_j \in TStest$  are considered.

Dissimilarity function for time series will be set as follows:

$$D(I_j) = \frac{1}{N} \cdot \sum_{a \in I_j} |a - \bar{I}_j|^2 \quad \text{where} \quad \bar{I}_j = \sum_{a \in I_j} \frac{a}{|I_j|}$$

First  $I$ , the average for the time series of  $I_j$  is calculated. Dissimilarity function is calculated as the sum of squared distances between the mean and vectors of  $I_j$ . The cardinality function is given by the formula  $C(I - I_j) = 1 / |I_j| + 1$ . The formula for calculating the smoothing factor is

$SF(I_j) = C(I - I_j) \cdot (D(I) - D(I_j))$ . If an exclusion-set  $I_x$ , received for  $I = TSStudy \cup \{ts\_test_j\}$  contains  $ts\_test_j$ , then  $ts\_test_j$  is an anomaly.

To determine anomalies in sets of time series based on the method described above the algorithm TS-ADEEP was developed.

In this paper, we propose the algorithm TS-ADEEP-Multi that is a generalization of the algorithm TS-ADEEP for the case of a learning set that contains examples of several time series classes. The generalization is quite obvious: dividing learning set into subsets containing examples of only one class and consistently applying to them and to each time series from a test set the algorithm TS-ADEEP, one can determine whether the considered time series are anomaly. For cases when

- 1) time series is an anomaly for each subset;
- 2) time series is not an anomaly for only a subset of the learning sample;

the answer is obvious. However, there is a case where the time series is not an anomaly for several classes of the learning set (but not all).

The pseudocode for algorithm TS-ADEEP-Multi is shown below:

The algorithm **TS-ADEEP-Multi**

input: (*TS Study*: learning set that contains examples of several classes;

*TSTest*: test set)

output: *TSAnom\_Optimistic* - a set of anomaly time series of on the "optimistic" assessment

*TSAnom\_Pessimistic* - a set of anomaly time series on the "pessimistic" assessment

**begin**

$TSAnom\_Optimistic = \emptyset$ ;  $TSAnom\_Pessimistic = \emptyset$

Let  $N$  be a number of classes containing in the learning set

$TSSStudy\_C = \{TSSStudy\_C_1, TSSStudy\_C_2, \dots, TSSStudy\_C_N\}$  is a partition of *TSSStudy* such that *TSSStudy\_C<sub>k</sub>* contains only examples of class  $k$ ,  $k = 1..N$

for  $j$  from 1 to  $|TSTest|$

    choose *TSTest<sub>j</sub>* of *TSTest*

    for  $k$  from 1 to  $N$

$I = TSSStudy\_C_k \cup TSTest_j$

        Find the exclusion-set  $I_x$  in  $I$

        If the  $TSTest_j \in I_x$  is an anomaly for the class  $k$  (it does not belong to him) then break

        If *TSTest<sub>j</sub>* does not belong to any of classes *TSSStudy\_C<sub>k</sub>*,  $k = 1..N$  then

$TSAnom\_Optimistic = TSAnom\_Optimistic \cup TSTest_j$

$TSAnom\_Pessimistic = TSAnom\_Pessimistic \cup TSTest_j$

        If *TSTest<sub>j</sub>* belongs to a unique class *TSSStudy\_C<sub>k</sub>* then it is not an anomaly

        If *TSTest<sub>j</sub>* belongs to several classes *TSSStudy\_C<sub>k</sub>* then

$TSAnom\_Pessimistic = TSAnom\_Pessimistic \cup TSTest_j$

print *TSAnom*

**end**

### **The simulation results for anomaly detection in time series**

A simulation of anomaly detection process was conducted on synthetic and real data. As the synthetic data were taken classic time series description used in the scientific literature: «cylinder-bell-funnel» [Naoki, 1994] and «control chart» [Pham and Chan, 1998]. As the real was used data collected through special systems traffic analysis during files transfer via various protocols.

«Cylinder-bell-funnel» [Naoki, 1994.] contains three different classes - "cylinder", "bell", "funnel".

«Control chart» [Pham and Chan, 1998] contains six different classes that describe the trends may be presented in the process: cyclical, decreasing value, sharp drop, increasing value a constant, sharp increase.

In order to determine how well the proposed algorithm deals with anomaly detection in time series, several experiments were conducted. Let consider the simulation data set «cylinder-bell-funnel». First, as a learning set  $TSS_{study}$  considers as a set of time series belonging to two of the three classes, for example, "cylinder" and "bell". As test set  $TSTEST$  considers as set of time series belonging to all three classes «cylinder», «bell», «funnel». Time series  $TStest_j$  is "normal" if it is a member of «cylinder» or «bell» classes and "abnormal" if it is not a member of them. Accordingly, the algorithm correctly finds anomalies, if it considers the time series of a class «funnel» from  $TSTEST$  as anomalies. It has been considered as a numerical representation of the time series and symbolic ones with a different alphabet sizes. Similarly, simulations carried out for the other pairs of classes: «bell» and «funnel», «cylinder» and «funnel».

The experiment showed that the proposed algorithm for anomaly detection does not always show good results: for some pairs of classes only a little more than half of the time series are correctly assigned to anomalies.

To improve this situation, it is proposed to further process the original time series by reduced or compressing it on normalization stage. This makes it possible to discard irrelevant details and to get rid from the noise. Example of time series compression is shown in Figure 3 (each ten points of original time series were assigned to a single point of new time series). The red lines in the figure connects points of original time series. The green line shows compressed time series: horizontal segments correspond to the 10 points of the original time series.

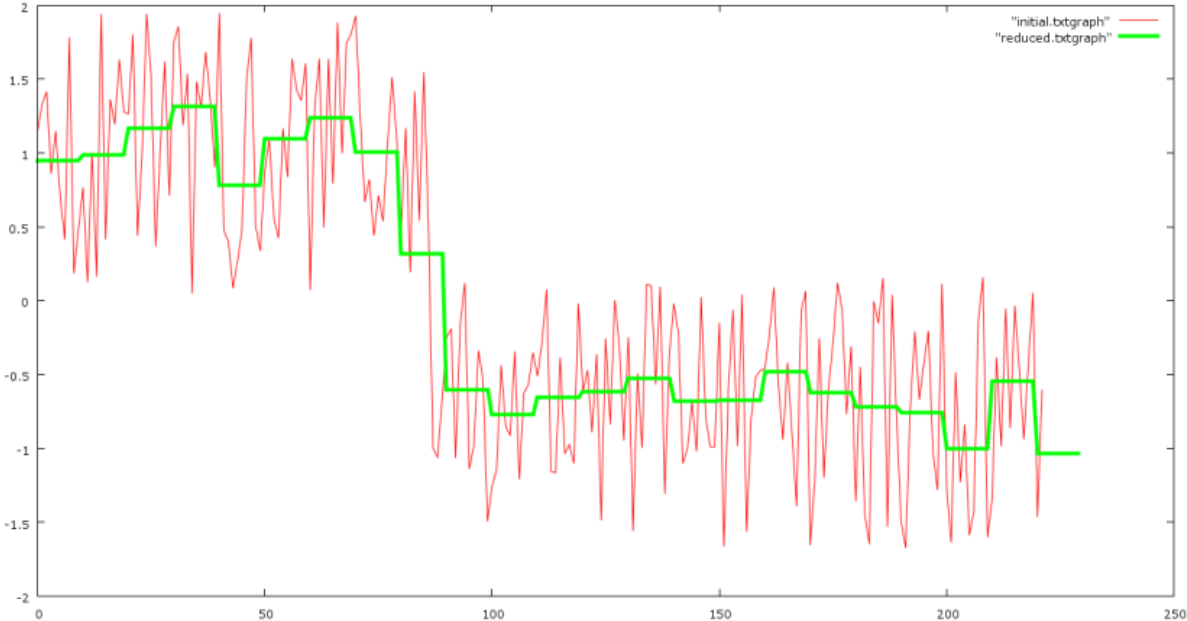


Figure 3. Example of time series compression



It has been verified in practice that the use of compressed time series algorithm for anomaly detection is significantly better than the use of it on the original data without compression (normal). Figure 4 shows a results comparison of a successful anomaly detection when using raw data without compression and compressed ones for «cylinder-bell-funnel». The Y-axis shows the percent of correctly recognized time series (normal or anomaly). The X axis presents four cases: a numerical representation of a value and a symbolic representation of a value (alphabet size 20, 35 or 50).

Figures 5 and 6 show results when using raw data without compression, and compression for the data sets «control chart».

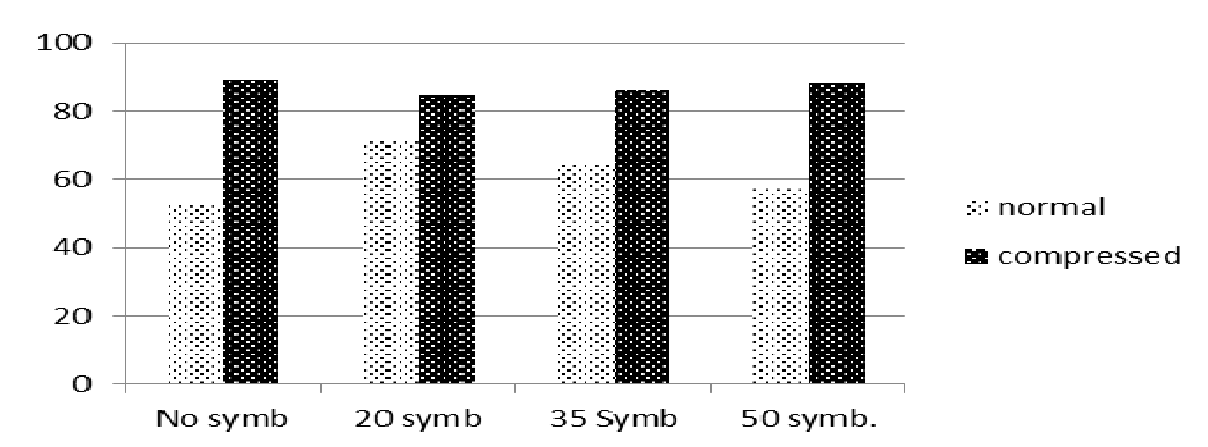


Figure 4. Comparison of a successful anomaly detection when using raw data without compression and compressed ones for «cylinder-bell-funnel»

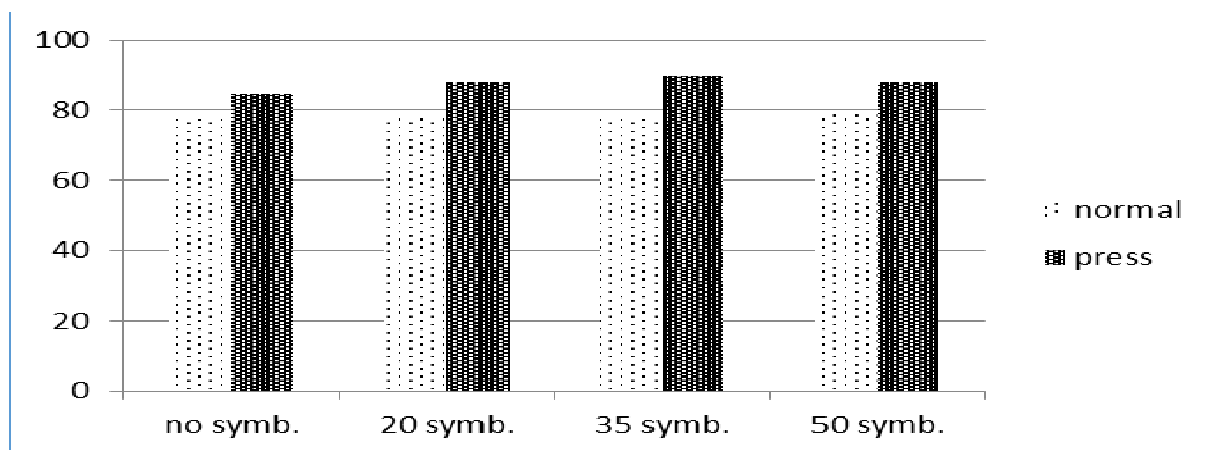


Figure 5. Comparison of a successful anomaly detection when using raw data without compression and compressed ones for «control chart», case of two classes

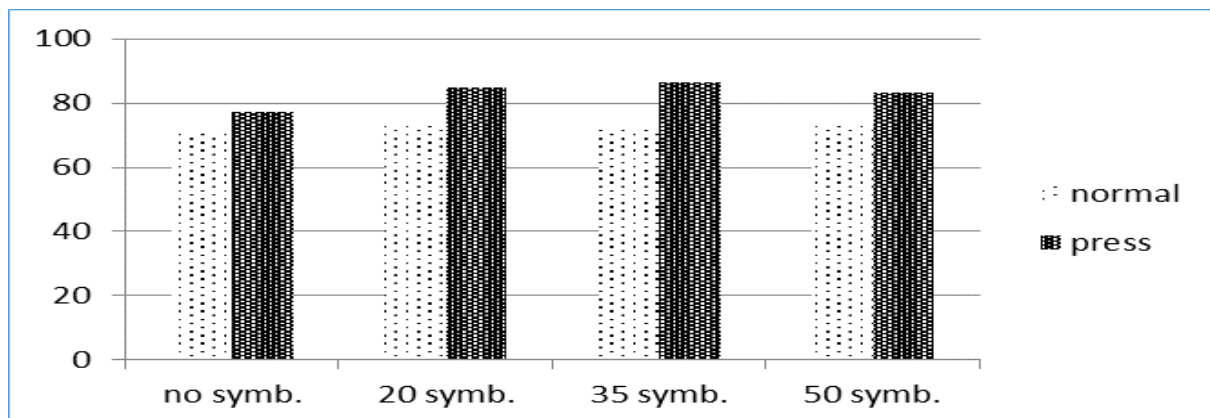


Figure 6. Comparison of a successful anomaly detection when using raw data without compression and compressed ones for «control chart», case of three classes

### Temporal Data Model

One of important problems arising under treating temporal data is a problem of representing temporal data in a convenient form under processing. There is supposed a temporal data model that can be used in intelligent decision support systems for storing and converting temporal dependencies [Eremeyev et al., 2010].

The object  $O$  of a temporal data model (TDM) is any information or structural entity, for example, domain, attribute, relation defined in a time interval  $I$ . In TDM a structure or values of any object are changed in time (what is typical for time series), therefore a category of time is a basic entity. The operator of a range of the definition for  $O$  is  $T: O \rightarrow I$  that returns a time interval when an object was defined. Let's consider TDM in a close analogy with a relation model since basic entities of both models coincide.

It is known that in a relation model, a domain is a set of one-typed values, for example, an integer set. A domain is simple if all its values are atomic (undecomposed) [Connolly, 2002]. A domain in TDM also satisfies these requirements, however it is defined in some time interval. A temporal cortege is determined in an interval  $I$  if in  $I$  domains of all attributes and its values are known in any moment from  $I$ .

In a relational model, a relation consists of cortege sets and each cortege has the same attribute set. For a temporal relation, a range of the definition (life time) is given through ranges of all corteges (records) inputting in a relation. In the general case, a cortege structure of a temporal relation can be arbitrary changed. In this connection, the identification problem of corteges appears. There is introduced an operator that for each cortege defines some unique key that is not changed under changing a cortege structure or values of cortege fields. A set of possible keys is associated with every relation  $R$ .

A key, as any element of TDM, has an own range of the definition, therefore, for choosing a primary key from the whole set of possible keys, it needs to choose a key that is defined in the whole required interval.

**Constrains of TMD integrity.** Rules of TMD integrity are intended for verifying entities and reference integrity. Integrity of entities means that at each moment, a value of primary key is one-valued determined, and reference integrity means that for each value of an external key appearing in a sibling relation, it needs to find a cortege with the same value of a primary key in a parent relation. And a definition range for values of an external key should be included in a definition range of a primary key.

Let's consider data processing in TMD. For this purpose, the operations for manipulating n-ary temporal relations are defined. Corresponding operations for a relational model are described in [Connolly, 2003]. Further  $R$  and  $S$  – relations;  $A, B, C$ , and so on (possible with indexes) – collections of attributes;  $c$  – a cortege of corresponding degree with corresponding domains.

The operation *THETA-SELECT* (*constrain*). The result of performing the operation:

$R[A \theta c]$  – a cortege set from  $R$ , each of which satisfies a condition that  $A$  – component is in the relation with the  $B$  – component. If the relation  $\theta$  is equality (widespread case), the operation *THETA-SELECT* is called simple *SELECT*.

The operation *PROJECTION*. The result of performing the operation:  $R[A_1, A_2, \dots, A_n]$  – the relation obtained by deleting all columns from  $R$  with the exception that are specified by attributes  $A_1, A_2, \dots, A_n$  and following deleting surplus line - duplicates and a range of definition projection coincides with a definition range of an original relation.

Operation *THETA-JOIN*. The result is a concatenation of relation lines  $R$  with relation lines  $S$  in accordance with the given condition defined by the relation  $\theta$ . :For TDM, a range of definition *THETA-JOIN* correspond to an interval where the relation  $\theta$  has been performed. If the relation  $\theta$  is equality then the operation *THETA-JOIN* is called *EQUI-JOIN*.

Operation *NATURAL JOIN*. This operation is analogous to the operation *EQUI-JOIN* with the exception that in this case surplus columns generated under performing join are eliminated. Natural join is join used under normalization of relation collection.

Operation *DIVIDE*. Let relation  $R(A, B_1)$  and  $S(B_2, C)$  be given such that  $B_1$  and  $B_2$  are defined on the same domain ( $S$ ). Then the result of this operation is a maximal subset  $R[A]$  such that its Cartesian product with  $S[B_2]$  is included into  $R$ .

The possibility of a TDM implementation as a dialect of the widespread SQL on the basis of often used and capable of adaptation of DBMS with the open code SQLite has been considered in [Eremeev et al., 2012]. The given model is a base for realization of the temporal DB of an intelligent decision support system (real time) and allows to operate with temporal dependencies including time series.

---

## Conclusion

---

In this paper two approaches to processing of temporal data are considered. The approach based on clustering was applied to the solution of the problem of speaker clustering. Mel-frequency cepstral coefficients were used as speaker features. We propose a use of self-organizing incremental neural networks, because they have an ability for life leaning and no need in a priori knowledge about speakers or their quantity. A use of extended feature vectors with MFCC dynamics can improve the accuracy of speaker clustering. Currently the possibility of segmentation of a recording unit is under study for recordings that contain voice of two or more speakers.

Next we consider the problem of anomaly detection among sets of time series. We propose a nonparametric algorithm TS-ADEEP-Multi for anomaly detection in time series sets for the case when the learning set contains examples of several classes. The method for improving the accuracy of anomaly detection, due to "compression" of these time series is used to get rid of unnecessary detail and noise. In the future it is expected to modify the proposed algorithm to define abnormalities in the sets of time series for the case when the classes of time series are not known a priori.

---

## Acknowledgements

---

The work is executed at financial support of RFBR (projects 14-07-00862, 14-01-00427, 15-01-05567) and in frame of Scientific research work T 2.737.2014/K of the project part of the government job in the sphere of scientific activity.

---

## Bibliography

---

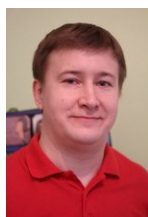
- [Abdallah et al., 2012] S.J. Abdallah, I.N. Osman, M.E. Mustafa Text-Independent Speaker Identification Using Hidden Markov Model In: World of Computer Science and Information Technology Journal (WCSIT), 2012, pp. 203-208.
- [Agrawal and Srikant, 1995] R. Agrawal, R. Srikant Mining Sequential Patterns In: Proceedings of the Eleventh International Conference on Data Engineering. ICDE '95. Washington, DC, USA: IEEE Computer Society, 1995, pp. 3-14 .
- [Ajmera and Wooters, 2003] J. Ajmera, C. Wooters A Robust Speaker Clustering Algorithm In: IEEE Workshop on Automatic Speech Recognition and Understanding, 2003, pp. 411-416.
- [Arning et al., 1996] A. Arning, R. Agrawal, P. Raghavan A Linear Method for Deviation Detection in Large Databases In: Proceedings of KDD'1996. 1996. pp. 164–169.
- [Antunes and Oliveira, 2001] M. Antunes, A. L. Oliveira Temporal data mining: an overview In: Eleventh International Workshop on the Principles of Diagnosis. 2001.

- [Boersma and van Heuven, 2004] P. Boersma, V. van Heuven Praat, a system for doing phonetics by computer In: Glot International 5(9/10), 2004, pp. 341-345.
- [CMU Arctic] CMU Arctic: [http://www.festvox.org/cmu\\_arctic/](http://www.festvox.org/cmu_arctic/)
- [CMU Chaplain] CMU Chaplain: <http://www.speech.cs.cmu.edu/Tongues/>
- [CMU Sphinx] CMU Sphinx: <http://cmusphinx.sourceforge.net/wiki/>
- [Connolly and Begg, 2002] T.M. Connolly, C.E. Begg Database Systems: A Practical Approach to Design, Implementation, and Management. Third Edition, Addison-Wesley, 2002, 1324 p.
- [Cooley and Tukey, 1965] J.W. Cooley, J.W. Tukey An Algorithm for the Machine Calculation of Complex Fourier Series In: Mathematics of Computation, 1965, pp. 297-301.
- [Eremeev et al., 2010] A.A. Eremeev, A.P. Eremeev, A.A. Panteleev Temporal data model and possibilities of its implementation on the basis of technology OLAP In: XII National Conference on Artificial Intelligence with international participation (CAI-2010):Proceedings in 4 volumns. Vol.3. – M.:Fizmatlit, 2010, pp.345-353 (in Russian).
- [Eremeev and Panteleev, 2012] A.P. Eremeev, A.A. Panteleev About possibility of implementation of the temporal query languages In: J. MPEI Bulletin, № 2, 2012, pp.155-160 (In Russian).
- [Furao and Hasegawa, 2006] S. Furao, O. Hasegawa: An incremental network for on-line unsupervised classification and topology learning // Neural Networks Vol. 19 Issue 1, 2006, pp. 90-106.
- [Furao et al., 2007] S. Furao, T. Ogura, O. Hasegawa An enhanced self-organizing incremental neural network for online unsupervised learning In: Neural Networks Vol. 20 Issue 8, 2007, pp. 893-903.
- [Han and Narayanan, 2008] K.J. Han, S.S. Narayanan Agglomerative Hierarchical Speaker Clustering using Incremental Gaussian Mixture Cluster Modeling In: Proceedings of InterSpeech, 2008, pp. 20-23.
- [Kumar and Rao, 2011] Ch.S. Kumar, P. M. Rao: Design of an Automatic Speaker Recognition System using MFCC, Vector Quantization and LBG Algorithm In. International Journal on Computer Science and Engineering (IJCSSE) Vol. 3 Issue 8, 2011, pp. 2942-2954.
- [Lin et al., 2003] J. Lin, E. Keogh, S. Lonardi and B. Chiu A Symbolic Representation of Time Series, with Implications for Streaming Algorithms In: Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2003, pp. 2-11.
- [Linde et al., 1980] Y. Linde, A. Buzo, R.M. Gray An Algorithm for Vector Quantizer Design In: IEEE Transactions on Communications vol.28 iss.1, 1980, pp. 84-95.
- [Molau et al., 2001] S. Molau, M. Pitz, R. Schlüter, H. Ney: Computing mel-frequency cepstral coefficients on the power spectrum In IEEE International Conference on Accoustics, Speech, and Signal Processing Vol.1, 2001, pp.73-76.

- [Mori and Nakagawa, 2001] K. Mori, S. Nakagawa Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition In: Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on Vol.1 , 2001, pp. 413-416.
- [Ning et al., 2006] H. Ning, M. Liu, H. Tang, Th. Huang: A Spectral Clustering Approach to Speaker Diarization In: Proceedings ICSLP, 2006.
- [Perfilieva et al., 2013] L. Perfilieva, N. Yarushkina, T. Afanasieva, A. Romanov Time series Analysis using Soft Computing Methods In: International Journal of General Systems, 42(6), 2013, pp.687-705.
- [Pham and Chan, 1998] D.T. Pham, A.B. Chan Control Chart Pattern Recognition using a New Type of Self Organizing Neural Network In: Proceedings. Instn, Mech, Engrs. Vol 212, No 1, 1998, pp. 115-127.
- [Rabiner, 1989] L. Rabiner: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition In: Proceedings of the IEEE vol.77 iss.2, 1989, pp. 257-286.
- [Roddick and Spiliopoulou, 1999] J.F. Roddick, M. Spiliopoulou A bibliography of temporal, spatial and spatio-temporal data mining research In: SIGKDD Explor. Newsl. 1999, jun., Vol. 1, No. 1. pp. 34–38.
- [Roweis, 1998] S. Roweis: EM Algorithms for PCA and SPCA In: Advances in Neural Information Processing Systems, 1998, pp. 626-632
- [Saito, 1994] N. Saito Local feature extraction and its application using a library of basesn PhD thesis, Yale University, December 1994.
- [Stefano et al., 2000] C. De Stefano, C. Sansone, M. Vento To reject or not to reject: that is the question — an answer in case of neural classifiers In: IEEE Transactions on Systems, Management and Cybernetics Vol. 1, 2000. pp. 84–94.
- [Vagin et al., 2008] V. Vagin, E. Golovina, A. Zagoryanskaya, M. Fomina Exact and Plausible Inference in Intelligent Systems 2-nd Edition, 2008 (in Russian).
- [Vyas and Kumari, 2013] G. Vyas, B. Kumari Speaker Recognition System Based on MFCC and DCT In: Internatonal Journal of Engineering and Advanced Technology(IJEAT) Vol. 2, Issue 5, 2013.
- [Weiqiang et al., 2002] W.Lin, M.A. Orgun, G.J. Williams An Overview of Temporal Data Mining In: Proceedings of the 1st Australasian Data Mining Workshop, 2002.
- [Zhu et al., 2005] X. Zhu, Cl. Barras, S. Meignier, J.-L. Gauvain Combining Speaker Identification and BIC for Speaker Diarization In: Interspeech'05 ISCA, 2005

## Authors' Information

---



**Antipov Sergey** – M.Sc., postgraduate student, Applied Mathematics Department of National Research University "Moscow Power Engineering Institute", 14, Krasnokazarmennaya Str., Moscow, 111250, Russia, Moscow, e-mail: [antisergey@gmail.com](mailto:antisergey@gmail.com)

**Area of scientific interests:** artificial intelligence, time series processing, decision support system.



**Fomina Marina** – Associated Professor, Computer Science Department of National Research University "Moscow Power Engineering Institute", 14, Krasnokazarmennaya Str., Moscow, 111250, Russia, Moscow, e-mail: [m\\_fomina2000@mail.ru](mailto:m_fomina2000@mail.ru)

**Area of scientific interests:** Inductive notion formation, Knowledge discovery in Databases.



**Vagin Vadim** – Ph.D., Professor, Applied Mathematics Department of National Research University "Moscow Power Engineering Institute", 14, Krasnokazarmennaya Str., Moscow, 111250, Russia, Moscow, e-mail: [vagin@appmat.ru](mailto:vagin@appmat.ru)

**Area of scientific interests:** artificial intelligence, logics, expert system.



**Ereemeev Aleksandr** – Ph.D., Professor, Head of the Applied Mathematics Department of National Research University "Moscow Power Engineering Institute", 14, Krasnokazarmennaya Str., Moscow, 111250, Russia, Moscow, e-mail: [eremeev@appmat.ru](mailto:eremeev@appmat.ru)

**Area of scientific interests:** artificial intelligence, decision making, decision support system, expert system.



**Ganishev Vasilii** – M.Sc., postgraduate student, Computer Science and Embedded Systems, Ilmenau University of Technology, 5, Helmholtzplatz, Ilmenau, 98693, Germany, e-mail: [vasily.ganishev@tu-ilmenau.de](mailto:vasily.ganishev@tu-ilmenau.de)

**Major Fields of Scientific Research:** process mining, workflow management, artificial intelligence, neural networks.

## INFORMATION FLOWS ENHANCEMENT FOR AIS TOURISM AUTOMATED INFORMATION SYSTEM

Irina Titova, Natalia Frolova

**Abstract:** *This article put emphasis on the issue of information system efficiency in tourism, as a tool of defining the level of domestic and inbound tourism growth. The necessity of information system development is determined to solve the problem of current tourism functionality; however, the implementation of some automated information system (AIS) cannot be a solution to the obstacle. AIS need to be efficient. Thus, the article provides the way of estimation and enhancement of information flows as an approach to improve information system efficiency.*

**Keywords:** *automated information systems (AIS), efficiency, information flow, graph modeling, AIS enhancement, information systems in tourism.*

**ACM Classification Keywords:** *G.2.2 Graph Theory, H.1 Models and Principles, K.4.3 Organizational Impacts*

---

### Introduction

---

The Russian Federation (RF) has a high tourist and recreational potential. There are a variety of cultural, historical and entertaining areas here. The immense territory rich in unique natural attractions, but its potential is realized incompletely.

In order to change this situation, a Federal target program for domestic and inbound tourism to 2020 was developed (FTP) [Collection of Legislation, 2011]. This program aims to increase the competitiveness of the tourism market of the Russian Federation, by providing high-quality tourism services. The goal is going to be achieved through the creation of conditions for tourism activities by:

- development of tourist and recreational complex of the Russian Federation;
- improvement of the quality of tourism services;
- advancement of national tourism product for domestic and international public.

Specification of the tourism industry realization ways improvement is made by extracting the key areas of support and development of the industry, by determining the methods of tourism enhancement in the country, and also by defining target indicators of FTP efficiency assessment. Tourist conferences, forums and exhibitions have become global events, with the honored participants, including government



officials. The main problems noted during the meetings are the adaptation of domestic tourism to the new realities and prospects of development in new conditions.

International events organized in Russia such as the Winter Olympics in Sochi, the World Championships in Aquatics in Kazan, the World Cup, Summits SCO and BRICS in Ufa, East Economic Forum in Vladivostok, stimulate facilities creation, which has a beneficial effect on the development of domestic and inbound tourism. The current socio-economical and political situation even in greater extent demonstrates the necessity of tourism development in Russia.

development of the industry is becoming based on the FTP. It includes the establishment of information exchange between Federal Agency for Tourism and regional departments. Automated information system (AIS) is one of the most progressive areas and a tool of information support for FTP tasks. AIS takes a role of means of storing, processing data, its visualization and dissemination.

The development of the information system aimed to provide information accessibility for different user groups, including executive personnel, involves the electronic format of the interaction between employees of federal and regional levels. Implementation of the information system is a decision of problems of information availability, reliability and timeliness. Thus, information is essential organization resource.

Information acts as a source for decision-making, because it integrates experience of routine activities.

Providing organization with urgent information is important to decision making quality, in order to make this process and the quality of the decisions worthy and to meet the requirements of current conditions [Shilyaev, 2005]. In other words, information quality should not prevent normal organization work, and, if possible, it should stimulate further development.

Special attention should be given to the question of the information quality estimation especially in the context of the existing information problems of public tourism organizations such as the lack of information unification, its fragmentation and territorial remoteness of interacting organizations. Information - data organized in a manner that it makes sense for a person. In this system, information - some of the key data which is used in reports.

The transformation process from the original row data to the end report through intermediate indicators is determined by information flows. Information flows - are an organized set of interrelated information blocks, which together provide a valuable result for customers. In this subject area information flows take into account the movement of raw data rather than the change of final documents. A valuable result for customers are based on a set of input data, such as agreements, contracts, acts or bank orders.

Z.V. Alferova and V.P. Ezzhaeva note the need for a long and careful analysis of the automation object as the part of information flows structure and information processing. The authors suggest as a research

approach methods based on graph theory because graphs allow to determine correlation between initial and final data [Alferova, 1971].

Implementation of specialized information systems is not a solution to the problem of information support, it is just a tool and technology for management personnel [Belyaev, 2010]. Information system has become a fundamental factor in the development in the case when it is effective.

There are many approaches to determine efficiency. Classical approach to the assessment considers the overall efficiency from the standpoint of three components: institutional, social and economic efficiency [Belyaev, 2010]. Each component is characterized by its objectives: the solution of problems, meet the expectations of employees and the achievement of financial results.

Other researchers consider methods for estimating benefits from the information systems implementation, focusing on two sides - tangible and intangible. First means measurable quantitative indicators (including financial results), second characterizes complex quality indicators (such as loyalty, performance, satisfaction) [Serdenko, 2014]. Thus, organization or information system can be specified as a subject of research.

This article is devoted to the assessment of the information system itself, because its effectiveness directly determines the efficiency of the organization. By efficiency we mean the ability of the system during its operation to produce the effect. In other words, efficiency is the ability of the system to bring some benefit. The greater benefit it creates; the higher efficiency is. The concept of efficiency in this situation is closely correlated with the term "quality". System vulnerability demonstrates low quality of product and, consequently, its low efficiency.

### Approach

The tourism automated information system was inspected for efficiency. The problem of data actualization delay has been identified as a result of modeling, supervision, documenting and questioning. It associates with time separation of the data input stage and the conversation stage (Figure 1). The delay in form renewal determined by the duration of calculations.

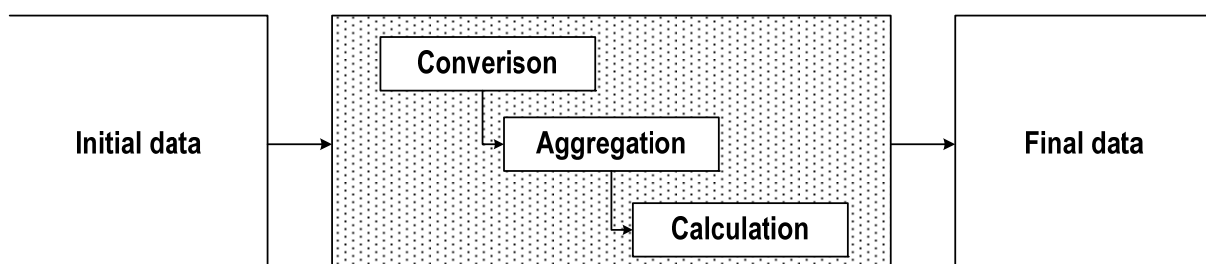


Figure 1. Data transformation

Thus, the process of collecting and reporting data is critical in the system. This process should be analyzed more carefully. A set of initial data becomes a desired form by the processing. Data transformation can include three phases: conversion, aggregation and calculation. Aggregation and calculation stages can be omitted.

Conversion is a data transformation from initial data to some intermediate form (for example cube or report) due to applying filter conditions. Aggregation is a data summation on the various report's levels (for example, aggregation of regional data to the RF level). Calculation is a mechanism for obtaining new intermediate forms based on previously acquired data.

The data processing is a critical step (operation) for the system. Critical work in the system is defined closely to critical work in the project as a work at critical path. Critical path is a sequence of operations, which influence on the process duration and which define working schedule [Bowen, 1997]. This stage determines data actuality in the system so it is important to realize it effectively. Therefore, found bottleneck of information flow realization should be explored in detail.

Algorithm analysis involves the research of the information flow transformation. The transformation process can be represented by a graph model, implemented by semantic network or directed graph (Mor1). Moreover, this graph will be isomorphic to initial algorithm.

Information graph is a diagram of information flows of the modeled system. It reflects all sequences for this information flow from initial to final data. Using the graph allows you to visualize the structure of the information flow, to apply a standard set of operations defined on graphs.

The task of improving information flows is a decision of graph reduction problem. Graph reduction includes merging and splitting of vertices in a manner that algorithm execution logic is saved, but its complexity is changed.

Bound for the complexity of the graph  $\theta(G)$  is carried out by estimation the complexity of the algorithm realized by the graph. The complexity of the algorithm is understood as a quantitative calculation of algorithm elementary operations.

Thus, the challenge of Information flows enhancement is the transformation of original graph to transformed graph  $G \rightarrow G'$  in such way that the complexity of transformed graph is less than the original one  $\theta(G) < \theta(G')$ . If reduced (transformed) graph has the least possible complexity  $\theta(G') \rightarrow \min$ , then  $G'$  is an optimized graph  $G$ .

Information flow model (Figure 1) includes three stages of data processing. Because of this organization, various groups of vertices are required:

- Initial documents, input forms or other data that is entered in the system by users (agreements, bank orders and other).
- Filter conditions are logical expressions that define the rules of data transformation.

- Intermediate documents are some data views (cubes,forms) that contain data for the construction of the final document or calculation methods.
- Aggregation mechanism are the conditions for data aggregation levels (total values).
- Calculation formulas are processes that produce new intermediate or final documents.
- Final documents final reports and forms.

Thus, we will use a marked graph  $G(V, E)$  with a set of vertices  $V$  and a set  $E$  of edges.

$V \subseteq V_1 \cup V_2 \cup V_3 \cup V_4$ , where

- $V_1$ - set whose elements are the initial, intermediate and final documents (data);
- $V_2$ - set whose elements are the filter conditions;
- $V_3$  - set whose elements are the aggregation mechanisms;
- $V_4$  - set whose elements are data calculation formula.

$E$  - set of edges that define the connection of vertices.

Figure 2. shows the structure of information flow for one chief process implemented in automated information system “AIS Tourism”.

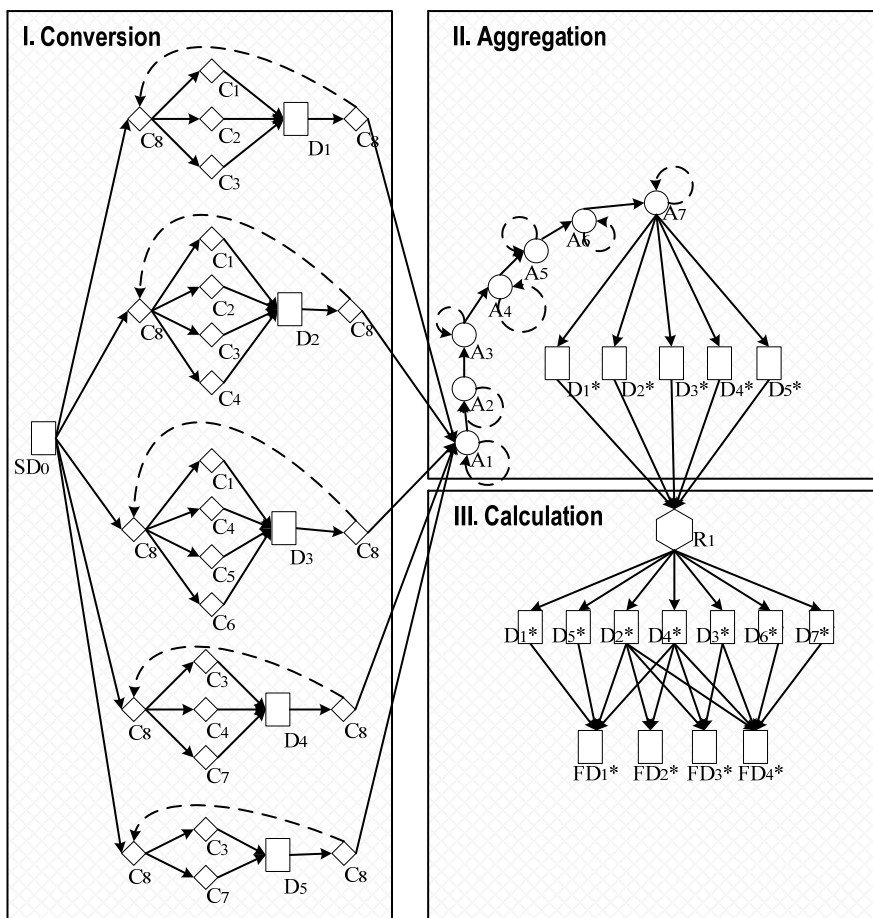






Figure 2. Initial information flow graph

For all vertices are defined symbols on the graph. The compliance of symbols and description is given in the Table 1.

Table 1. Graph description

Symbol	Description
	Initial (SD), intermediate (D) and final documents (FD)
	Filter conditions (C)
	Aggregation mechanisms (A)
	Calculation formulas (R)

This graph has multiple duplicated vertices which characterize additional complexity during the operations execution. Extra activities increase time of algorithm performance. In this way the growth of input data will increase the speed of algorithm execution because of its polynomial complexity. Graph complexity will be estimated by the amount of elementary operations or vertices.

AIS Tourism consists of  $n$  acts and bank orderd. Acts have just one third  $p_a \approx \frac{1}{3}n$  of all payment documents and bank orders have last two thirds  $p_{bo} \approx \frac{2}{3}n$ .

Filter conditions vertices are defined by amount of elementary operations. We will consider following operations as elementary:

- Logic operations (such as and, or, not, xor).
- Relational operators ( $=$ ,  $<>$ ,  $>$ ,  $<$ ,  $<=$ ,  $>=$ ).
- Integer arithmetic.
- Arithmetic ( $+$ ,  $-$ ,  $*$ ,  $/$ ).

The conversation part has the complexity (1), where  $k_i$  is amount of vertices numbered  $i$ ,  $a_i$  is the complexity of vertice numbered  $i$  and  $m$  is the highest vertice number of filter condition in studied group.

$$A = 5na_8 \sum_{i=1}^m k_i a_i \quad (1)$$

As a result, the complexity of the conversion part can be calculated by the formula (2).

$$A = na_8(a_1 + a_2 + a_3) + na_8(a_1 + a_2 + a_3 + a_4) + na_8(a_1 + a_4 + a_5 + a_6) + na_8(a_3 + a_4 + a_7) + na_8(a_3 + a_7)$$

Table 2 includes data that is necessary to estimate the complexity of the graph.

Table 2. Complexity of vertices

Vertice number	1	2	3	4	5	6	7	8
Complexity	1	3	5	2	1	1	1	1
Amount of vertices	3	2	4	3	1	1	2	5 <sup>1</sup>

The complexity of other stages (aggregation and calculation) we will equate as a constant  $T$ . Thus, the complexity of the whole graph will be defined as (3):

$$\theta(G) = A + T \quad (2)$$

Analyzing the graph it was found that there are duplications of vertices with redundant filter conditions, that undermine the effectiveness of the system:

- It increases process time.
- It reduces quality. At the time of performing a long calculation data loss is more likely.
- It reduces system reliability. Long consuming operation has a higher risk of failure;
- System seems to be more complicated.

Graph enhancement can be achieved due to reduction of graph based on merging and splitting of vertices. The complexity of operations (vertices) should be taken into account during the reduction process.

First thing that can be done is merging of vertices C8 that control loop operation. Initially there were five loops with the same terminate condition, so it can be simplified by removing needless one (Figure 3).

Figure 3.

<sup>1</sup> Two loop conditions C8 are determined by the same complexity, but on the graph it presents as double vertice.

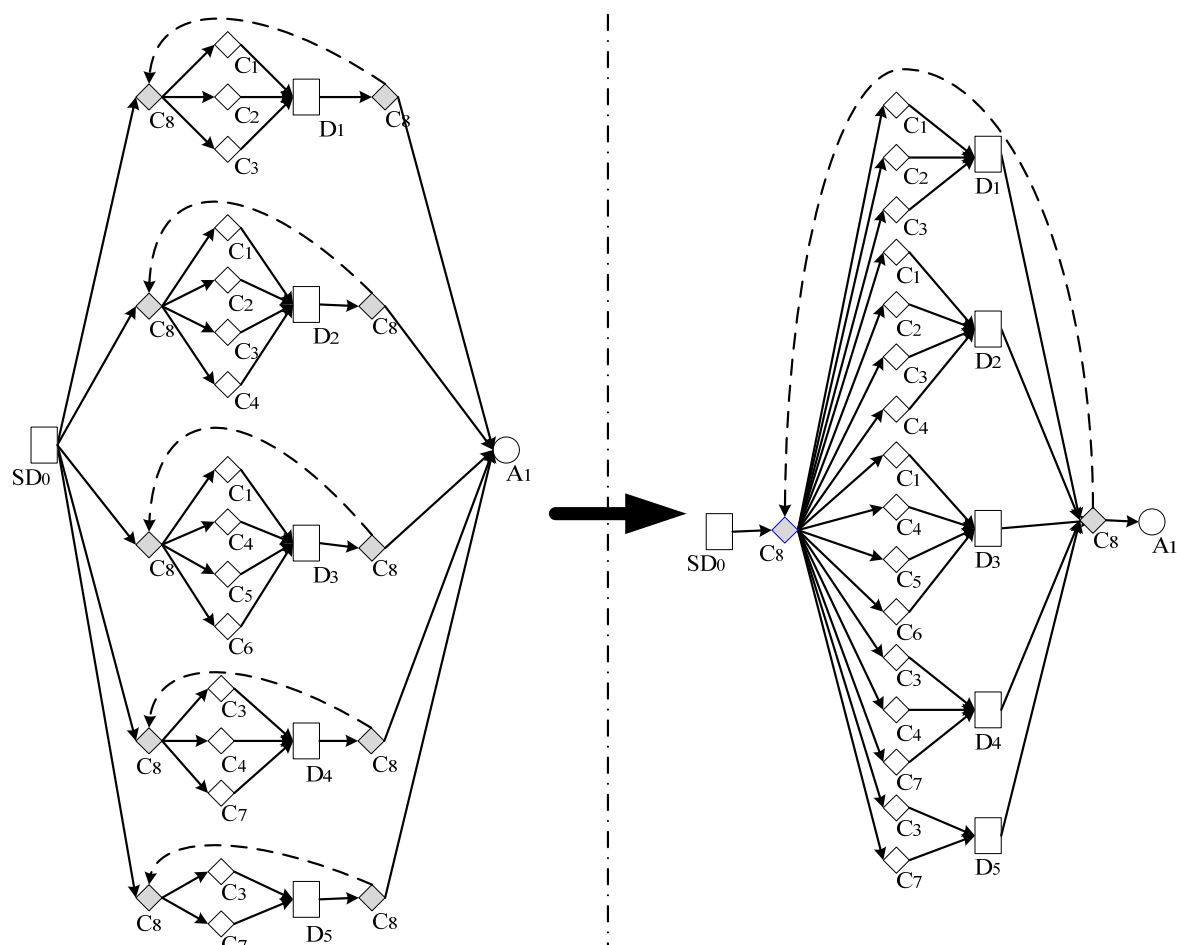


Figure 3. Transformation of vertex C8

Then there are conditions C1 and C7 that consist of opposite facts, it means that if C1 is true then C7 is false. The converse case is also true. So these vertices can be changed by the new vertex C9 with XOR mark. C9 will divide the data set into two groups (acts and bank orders) with success rates  $p_a$  and  $p_{bo}$ . Generation of new vertex require the creation of new intermediate documents SD1 and SD2 (Figure 4).

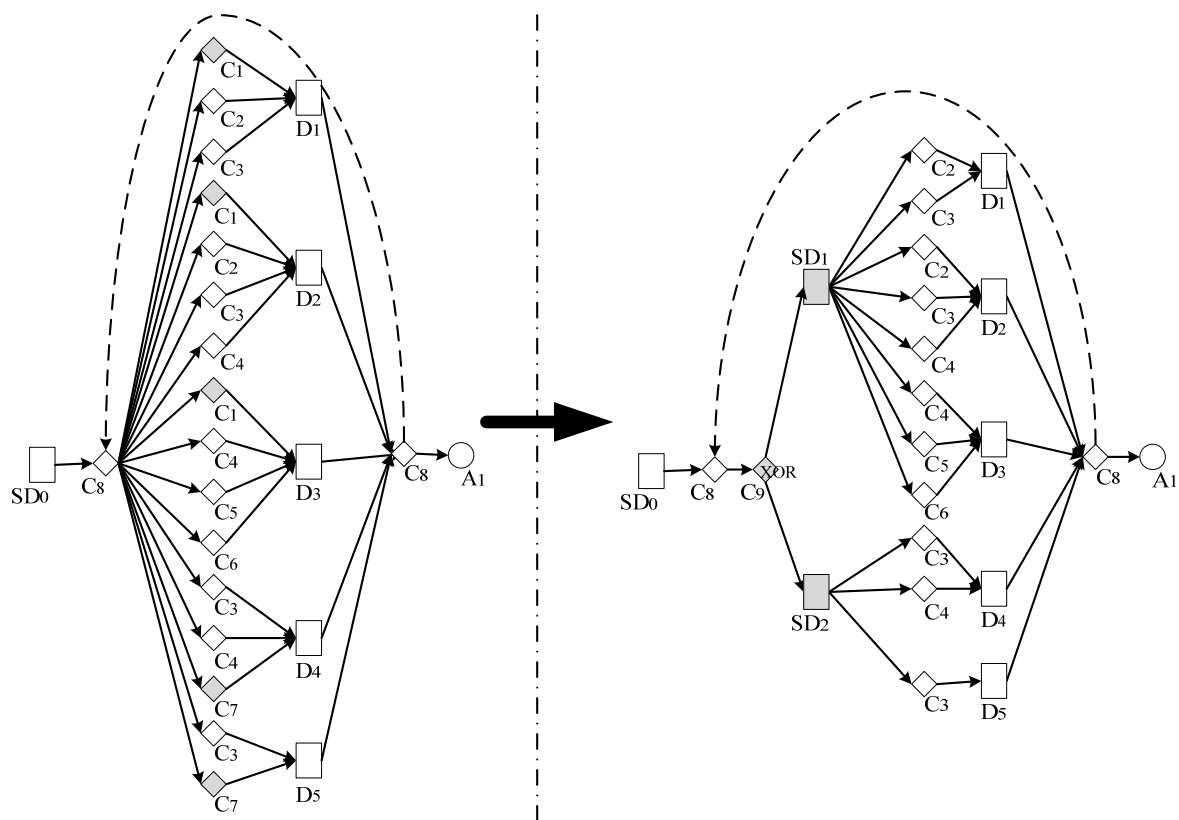


Figure 4. Generation of vertex C9

Each new group (characterized by acts and bank orders) can be implemented in individual reduced loop (Figure 5).

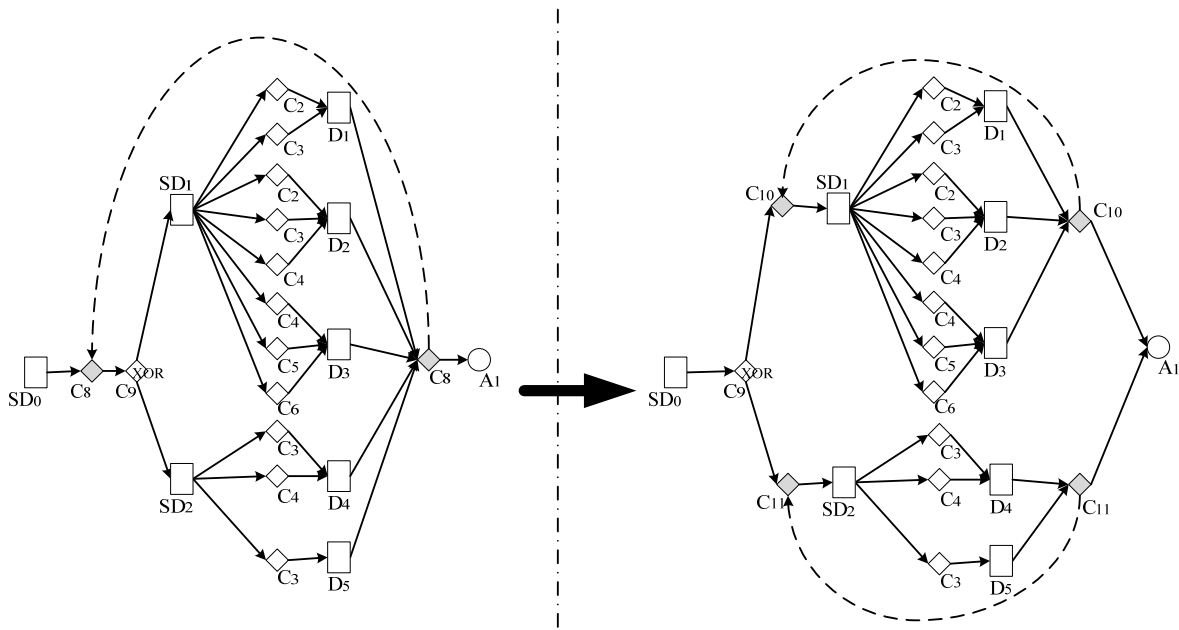


Figure 5. Simplification of C8 loop



After that vertices duplications can be excluded by merging of vertices C2 (Figure 6) and others.

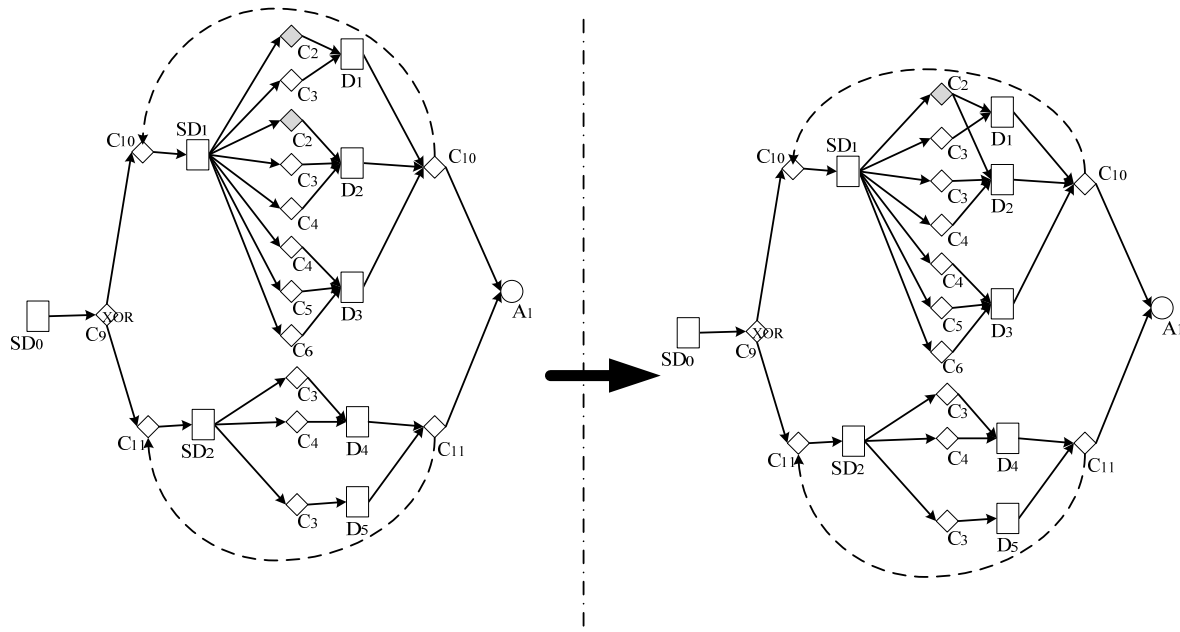


Figure 6. Merging of C2 vertices

As a result, it was received new transformed graph (Figure 7) with complexity (4).

$$\theta(G') = A' + T \tag{3}$$

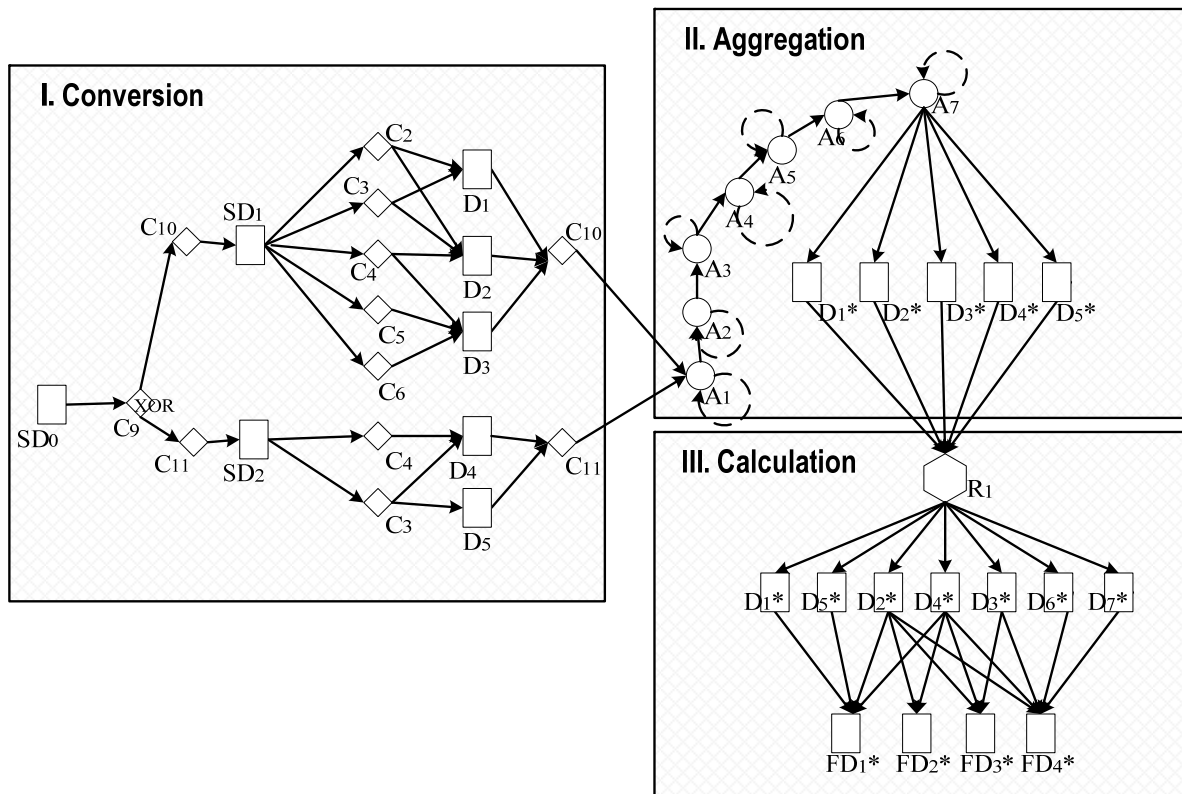


Figure 7. Transformed graph

The complexity of conversion stage of transformed graph  $G'$  can be calculated by formula (5).

$$A' = n \cdot p_{bo} \cdot a_{10} \cdot \sum_{i=2}^6 a_i + n \cdot p_a \cdot a_{11} \cdot (a_3 + a_4) \quad (4)$$

Therefore, the complexity of conversion part of transformed graph is less than original graph complexity  $A' < A$ , so total graph complexity of transformed graph is also less than original one  $\theta(G') < \theta(G)$ . Algorithm implemented by the information flow graph can be modernized by transformed graph.

---

## Conclusion

This research represents domain-specific method of estimation and enhancement of information flows for automated information system “AISTourism“. The method include three main parts: visualization, analysis or estimation and transformation.

Visualization includes graph modelling for previously defined vulnerabilities of effectiveness of the automated information system. Graph model implements the information flow. Graph realization was chosen due to its opportunity to use basic graph theory rules and its isomorphism to initial algorithm.

Evaluation of information flow model was carried out by applying the theory of algorithms. Improving the implementation of the information flow has been achieved through the transformation of the graph due to the merging and slitting operations.

---

## Bibliography

- [Alferova, 1971] Z.V. Alferova. Application of graph theory in economic calculations. Ed. Z.V. Alferova and V.P. Ezzhaeva. M.: Statistics, 1971. 150 p.
- [Belyaev, 2010] D.A. Belyaev. Fundamentals of Information Management: supportive notes of lectures. Syktyvkar: Syktyvkar University Publishing, 2010. 64 p.
- [Bowen, 1997] H.K. Bowen. Project Management Manual. Boston: HarvardBusiness School Press, 1997. 42 p.
- [Collection of Legislation, 2011] Collection of Russian Federation Legislation, 22.08.2011, No 34, item 4966.
- [Morozov, 1982] V.P. Morozov. Features of design of economic information processing systems based on ES EVM. M.: Finance and statistics, 1982. 150 p.
- [Serdenko, 2014] E.S. Seredenko. Cost-effectiveness analysis of information systems. Cand. econ. sci. diss. M., 2014.
- [Shilyaev, 2005] A.A. Shilyaev. Information support for the restructuring of the enterprise management system: Cand. econ. sci. diss. Moscow, 2005.

---

## Authors' Information

---



**Irina Titova** – Business Analyst at Prognoz IT-company, e-mail: [ir.okulova@gmail.com](mailto:ir.okulova@gmail.com).

*Major Fields of Scientific Research: information flows, graph modeling.*



**Natalia Frolova** – Perm State National Research University, Doctor of Physical and Mathematical sciences, Associate Professor, Department of Information Systems and Mathematical Methods in Economics, E-mail: : [nvf\\_psu@mail.ru](mailto:nvf_psu@mail.ru)

*Major Fields of Scientific Research: information flows, graph modeling, complex system modeling.*

## TABLE OF CONTENTS

<i>Implementing a Linear Function to Measure the Quality in Governments.</i>	
Alberto Arteta, Juan Castellanos, Yanjun Zhao, Danush K Wijekularathna .....	3
<i>Collective computation: Turning the underground into an ant nest</i>	
Clemencio Morales, Luis Fernando de Mingo .....	12
<i>Clustering Using Particle Swarm Optimization</i>	
Nuria Gómez Blas, Octavio López Tolic.....	24
<i>On a Public Key Encryption Algorithm Based on Permutation Polynomials and Performance Analyses</i>	
Gurgen Khachatryan Martun Karapetyan.....	34
<i>convexity related ISSUES for the set of hypergraphic sequences</i>	
Hasmik Sahakyan, Levon Aslanyan.....	39
<i>Representing Strategic Organizational Knowledge via Diagrams, Matrices And Ontologies</i>	
Dmitry Kudryavtsev, Anna Menshikova, Tatiana Gavrilova.....	48
<i>Methods and Algorithms of Time Series Processing in Intelligent Systems</i>	
Sergey G. Antipov , Marina V. Fomina, Vadim V.Vagin, Alexandr P. Ereemeev, Vasilii A. Ganishev .	67
<i>Information Flows Enhancement for AIS Tourism Automated Information System</i>	
Irina Titova, Natalia Frolova .....	88
<i>Table of contents</i> .....	100