Computer Science Journal of Moldova, vol.**24**, no.2(71), 2016

Estimation of Morphological Tables Using Text Analysis Results*

Illia Savchenko

Abstract

This paper proposes methods for obtaining input data, necessary for the modified morphological analysis method, from the text sources of data using text analysis tools. Several methods are described that are suitable for calculating initial estimates of alternatives and cross-consistency matrix values based on processing text fragments by rule-based categorization and sentiment analysis tools. A practical implementation of this tool set for assessing statements in news regarding Ukraine is considered.

Keywords: Morphological analysis method, text analysis, morphological table, foresight.

1 Introduction

The modified morphological analysis method (MMAM) is a powerful quality analysis tool for problems, where the objects are characterized by uncertainty, incompleteness, inaccuracy, fuzziness of information and thus have a large multitude of possible alternative configurations [1]. These objects can be described and studied by MMAM, allowing to make decisions regarding such objects.

Work with objects in MMAM [2], [3] requires an initial estimation of alternatives and the cross-consistency matrix values, which is usually done by expert evaluation. This approach is the most exact, however, it requires a lot of time and financial resources. Besides, the number

*This research was sponsored by the NATO Science for Peace and Security Programme under grant 984877

of questions for experts even for relatively small morphological tables may amount to hundreds or thousands, which is inadmissible.

As the object descriptions in the foresight process are commonly verbal, it is proposed to involve text analysis tools for obtaining input values. Such analysis tool may include rule-based categorization and sentiment analysis tools [4], [5].

The MMAM in foresight problems is the most efficient one for two types of tasks [2]:

1. The description of objects (processes, phenomena), that emerge multiple times in different configurations, to study and make decisions against the possible multitude of such objects as a whole.

2. The description of a state of a given complex system, that has uncertain characteristics, e.g. a future state of this system.

The application of text analysis tools for estimating morphological tables is possible in both cases. In the first case the input data can be acquired from news, messages, claims regarding the objects of the studied type. In the second case information comes in the form of predictions, statements, comments of experts regarding the given system.

2 Statement of the problem

We will assume that we already have a collection of texts regarding the object, that is a target for morphological study.

The MMAM needs two types of input estimates: the initial probabilities of the alternatives, and the values of cross-consistency matrix for pairs of alternatives. To have the capacity to obtain this data from the unstructured information fragments, first we have to introduce the rules for text analysis [5], which allow relating the text with a certain degree of confidence to categories that correspond to the alternatives of the morphological table parameters.

Let's designate $R_j^{(i)}(g_k)$ as the rule for calculating the degree of relation for the text fragment g_k to a category that corresponds to the

alternative $a_j^{(i)}$ of the morphological table. This set of rules is constructed for each alternative $a_j^{(i)}$ of each parameter F_i of the morphological table. These rules determine that the text fragment mentions an object of study with the characteristic, specified by parameter F_i, has a value that corresponds to alternative $a_j^{(i)}$.

Then, the following statement of problem can be written:

Given:

- a morphological table with N parameters $F_i, i \in \overline{1, N}$, each having a set of alternatives $a_j^{(i)}, j \in \overline{1, n_i}$;
- a collection of N_{text} text fragments $g_k, k \in \overline{1, N_{text}}$. It is considered to be already determined that each of the text fragments mentions an object of study;
- $R_j^{(i)}(g_k), i \in \overline{1, N}, j \in \overline{1, n_i}, k \in \overline{1, N_{text}}$ – the text analysis rules for calculating the degree of relation of the text fragment g_k to the category, that corresponds to the alternative $a_j^{(i)}$ of the morphological table.

Required:

- to calculate the initial estimates $p'^{(i)}_j$ for each alternative $a_j^{(i)}$;
- to calculate the cross-consistency matrix values $c_{i_1 j_1 i_2 j_2}$ for each pair of alternatives $a_{j_1}^{(i_1)}, a_{j_2}^{(i_2)}$.

3 Evaluating initial alternative values

Analyzing a large enough collection of texts, we can make conclusions regarding the distribution of probabilities between the alternatives for any morphological table parameter. But it is necessary to mention that text analysis tools give only a hypothesis regarding the relation of a text fragment to certain alternative $a_j^{(i)}$, with a confidence of $R_j^{(i)}(g_k)$. Thus a situation is possible, when a single text fragment is related to several alternatives simultaneously with different degrees of confidence.

Taking this into account, we proposed two methods of evaluating initial alternative values:

Method 1: taking a ratio of total degree of relation to category that corresponds to $a_j^{(i)}$, to the total degree of relation for all categories that correspond to alternatives of parameter F_i, for all text fragments:

$$p'^{(i)}_j = \frac{\sum_{k=1}^{N_{text}} R_j^{(i)}(g_k)}{\sum_{k=1}^{N_{text}} \sum_{j^*=1}^{n_i} R_{j^*}^{(i)}(g_k)}.$$

Method 2: taking a ratio of a number of text fragments, where the degree of relation to $a_j^{(i)}$ is maximal, to the total number of text fragments, where the degree of relation to at least one alternative of the parameter F_i is larger than zero.

$$p'^{(i)}_j = \frac{\left|\left\{g_k | R_j^{(i)}(g_k) = \max\limits_{j^* \in \overline{1,n_i}} (R_{j^*}^{(i)}(g_k)), R_j^{(i)}(g_k) > 0\right\}\right|}{\left|g_k | \exists j^* \in \overline{1,n_i} : R_{j^*}^{(i)}(g_k) > 0\right|},$$

where $|A|$ is the power of set A, i.e. the number of elements in it.

The choice of method depends on the specifics of the problem. If the alternatives and their corresponding text analysis rules are defined in such a way that most text fragments relate to a single alternative, then the first method is more reliable. If most text fragments relate to multiple alternatives with positive degrees, then the second method may be more correct.

4 Evaluating cross-consistency matrix values

The cross-consistency matrix establishes the dependencies between the alternatives of different parameters, i.e. the type of influence of choosing one alternative in a pair on the probability of choosing the other one. For practical purposes this value can be regarded as the correlation between the choice of alternatives in a pair for the object configuration. Thus the first method of evaluating cross-consistency matrix values is formed.

Method 1. Using correlation between the degrees of confidence of relating the text fragments to alternatives from a pair, that is connected by the corresponding cross-consistency matrix value.

$$c_{i_1 j_1 i_2 j_2} = r(R_{j_1}^{(i_1)}, R_{j_2}^{(i_2)}) =$$

$$= \frac{\sum_{k=1}^{N_{text}} \big(R_{j_1}^{(i_1)}(g_k) - \overline{R_{j_1}^{(i_1)}}\big)\big(R_{j_2}^{(i_2)}(g_k) - \overline{R_{j_2}^{(i_2)}}\big)}{\sqrt{\sum_{k=1}^{N_{text}} \big(R_{j_1}^{(i_1)}(g_k) - \overline{R_{j_1}^{(i_1)}}\big)^2 \sum_{k=1}^{N_{text}} \big(R_{j_2}^{(i_2)}(g_k) - \overline{R_{j_2}^{(i_2)}}\big)^2}},$$

where $\overline{R_{j_1}^{(i_1)}} = \big(\sum_{k=1}^{N_{text}} R_{j_1}^{(i_1)}(g_k)\big)/N_{text}$ is the mean degree $R_{j_1}^{(i_1)}(g_k)$ for all text fragments. In this method only the text fragments with non-zero degrees of confidence for at least one alternative of each parameter F_{i_1}, F_{i_2} are considered:

$$g_k \in \big\{g_k | \exists j_1^* : R_{j_1^*}^{(i_1)}(g_k) > 0 \wedge \exists j_2^* : R_{j_2^*}^{(i_2)}(g_k) > 0\big\}.$$

The second method is based on the fact that the pair of alternatives, that have higher values in the cross-consistency matrix, also has the higher probability of mention in a text fragment.

Method 2. Calculating a ratio between the number of text fragments, where both of the alternatives are mentioned and the number of text fragments, where at least one alternative of the pair is mentioned:

$$c_{i_1 j_1 i_2 j_2} = 1 - 2\frac{\big|g_k | R_{j_1}^{(i_1)}(g_k) > 0 \wedge R_{j_2}^{(i_2)}(g_k) > 0\big|}{\big|g_k | R_{j_1}^{(i_1)}(g_k) > 0 \vee R_{j_2}^{(i_2)}(g_k) > 0\big|}.$$

For evaluating cross-consistency matrix the second method is more reliable in cases where most text fragments have a well-defined relation to a specific alternative. On the contrary, if the text fragments have non-zero degrees of relation to several alternatives of one parameter, then the first method is preferable.

5 Practical application

To test the developed tool set, it was employed to estimate the alternatives of a morphological table for international statements regarding Ukraine in 2013. The morphological table is presented in Table 1.

Table 1. Morphological table for statements regarding Ukraine

Statements regarding Ukraine		
1. Speaker	2. Subject	3. Tone
EU	Elections	Positive
NATO	Eurointegration	Neutral
Russia	Economics	Negative
USA		

To conduct evaluation, the SAS Content Categorization Studio software was applied, using a collection of 3805 text fragments (news in 2013). Elements of the building blocks for SAS text analysis rules are presented in Figure 1.

Using the proposed tool set, the initial values for the alternatives of the morphological table were obtained (Table 2):

Table 2. The initial values for the alternatives, obtained by the analysis of text fragments

Statements regarding Ukraine		
1. Speaker	2. Subject	3. Tone
0.569	0.216	0.051
0.064	0.386	0.903
0.307	0.398	0.046
0.06		

Similarly the cross-consistency matrix was evaluated (Table 3).

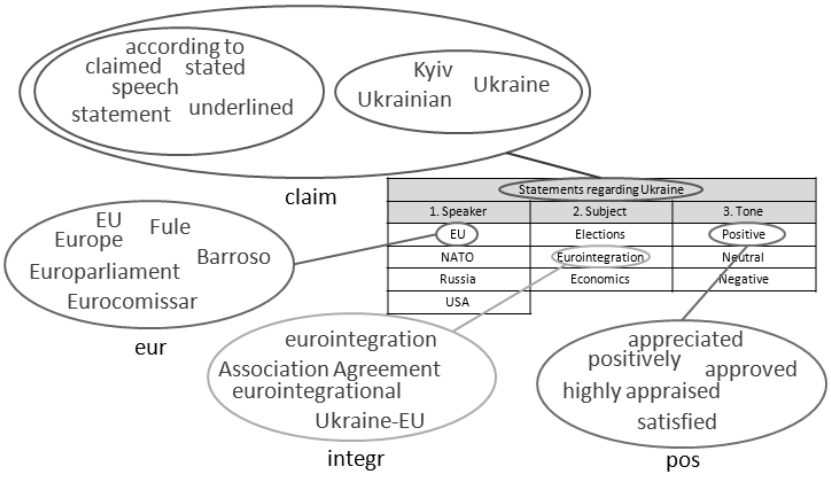

Figure 1. Samples of elements for text analysis rules

Table 3. The cross-consistency matrix values, obtained by the analysis of text fragments

		1. Speaker				2. Subject		
		$a_1^{(1)}$	$a_2^{(1)}$	$a_3^{(1)}$	$a_4^{(1)}$	$a_1^{(2)}$	$a_2^{(2)}$	$a_3^{(2)}$
2.Subject	$a_1^{(2)}$	–0.29	–0.27	–0.43	–0.13			
	$a_2^{(2)}$	0.25	–0.41	–0.39	–0.06			
	$a_3^{(2)}$	–0.32	–0.44	0.35	0.31			
3. Tone	$a_1^{(3)}$	–0.52	–0.71	–0.82	–0.78	–0.56	–0.16	–0.35
	$a_2^{(3)}$	0.59	0.51	0.76	0.38	0.87	0.87	0.9
	$a_3^{(3)}$	–0.52	–0.8	–0.88	–0.6	–0.51	–0.18	–0.39

6 Conclusions

The proposed above methods allow one to process large morphological tables without creating excess strain for experts. They also allow using in estimation the unstructured data, which is hard to account otherwise, to gather the statistics for the entities, where it would be inaccessible by other means.

The limitation of this method is the necessity of having a large enough volume of text information in the field of study. The input data must be processed as text fragments, each being a separate description of the object of study. The fragments also shouldn't duplicate the same description of the object, i.e. be independent of each other. A variety of thoughts and sources is encouraged to exclude one-sidedness and prejudice in estimates. Also the utilization of these methods require skills for creating and tuning text analysis tools.

Thus, the result of evaluation is highly dependent on the quality of text analysis rules and the collection of texts itself. These methods are rarely suitable as the only source of input data, however, they can be deemed as a starting point for further improvement, or as a thought of a single expert, when used in parallel with classic expert estimation.

References

[1] N. D. Pankratova, P. I. Bidyuk, Y. M. Selin, I. O. Savchenko, L. Y. Malafeeva, M. P. Makukha, V. V. Savastiyanov, "Foresight and Forecast for Prevention, Mitigation and Recovering after Social, Technical and Environmental Disasters," in *Improving Disaster Resilience and Mitigation – IT Means and Tools* (Part I), H.-N. Teodorescu, A. Kirschenbaum, S. Cojocaru and C. Bruderlein, Eds. Netherlands: Springer, 2014, pp. 119–134. Available: DOI: 10.1007/978-94-017-9136-6_8.

[2] N.D. Pankratova, I.O. Savchenko, *Morphological Analysis. Problems, Theory, Applications,* Kyiv, Ukraine: Naukova Dumka, 2015, 245 p. (in Ukranian)

[3] I.O. Savchenko, *Methodological and Mathematical Support for Solving Foresight Problems Using Modified Morphological Analysis Method.* Innovative Development of Socio-Economic Systems Based on Foresight and Cognitive Modeling Methodologies, Kyiv, Naukova Dumka, 2015, pp. 427–441.

I. O. Savchenko, "Methodological and Mathematical Support for Solving Foresight Problems Using Modified Morphological Analysis Method," in *Innovative Development of Socio-Economic Systems Based on Foresight and Cognitive Modeling Methodologies*, G. V. Gorelova and N. D. Pankratova, Eds. Kyiv, Ukraine: Naukova Dumka, 2015, pp. 427–441. (In Russian)

[4] B. Liu, *Sentiment Analysis and Opinion Mining,* Morgan & Claypool Publishers, 2012, 180 p. ISBN-13: 978-1608458844.

[5] H. Reckman, Ch. Baird, J. Crawford, R. Crowell, L. Micciulla, S. Sethi, F. Veress, "Rule-based detection of sentiment phrases using SAS Sentiment Analysis," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), Association for Computational Linguistics,* Atlanta, Georgia, 2013, pp. 513–519.

Illia Savchenko

Received February 20, 2016

Educational-Scientific Complex "Institute for Applied System Analyses",
National Technical University of Ukraine "Kyiv Polytechnic Institute"
prosp. Peremohy, 37, bd. 35, Kyiv, Ukraine.
Phone: +38 050 3871688
E–mail: savil@inbox.ru

Computer Science Journal of Moldova, vol.**24**, no.2(71), 2016

Tools for Texts Monitoring and Analysis Aimed at the Field of Social Disasters, Catastrophes, and Terrorism *

Svetlana Cojocaru, Mircea Petic and Grigorii Horos

Abstract

Throughout its life mankind faces various disasters and catastrophes: natural, technogenic, social. One of the important sources of information for these situations is the huge volume of unstructured data available in the global information networks. In this study, we describe a tool set that includes methods for extracting relevant texts from the networks, their classification and analysis. Two stages are described: preparatory and processing. The first one deals with patterns (texts and keywords) creation, during the second phase news texts are processed using database and controlled vocabulary.

Keywords: computational linguistic resources, linguistic markers.

1 Introduction

We live in an era that is increasingly affected by various natural disasters (earthquakes, forest fires, floods), together with technogenic disasters (explosions, leakage of toxic substances) or caused by individuals, such as terrorism.

A characteristic feature of contemporary society is the special role of information networks, where different signals related to disasters that may be produced or already have been produced, may occur promptly.

* This research is performed in the frame of a grant (SPS G4877) supported by the NATO international Science for Peace and Security Program.

This source of information can be used to prevent and mitigate the social consequences of disasters. It consists from a large volume of data, accessible on global information networks: mass-media, social networks, blogs, etc. A social disaster is usually followed by a huge amount of data generated in the form of news, discussion, expression of views etc. This information is quite difficult to process due to its unstructured form. In order to make right decisions in appropriate time, special analytical tools are needed that would give decision-makers support for a second opinion. Currently, such means practically do not exist, with the exception of guidance from a narrow range of applications (e.g., forest fire monitoring and forecasting). Achieving their implementation could significantly improve modelling and social disaster mitigation.

Our goal is to elaborate a tool which will collect and classify tweets and news texts related to disasters topics.

This article is carried out within a project [1] that aims creating a set of tools, methods and algorithms to detect different nature of social disasters. The main idea is to monitor information sources from network in three neighboring countries (Ukraine, Romania and Republic of Moldova), to find pertinent texts in four languages: Ukrainian, Russian, Romanian and English, to process them at Situation Analytical Center established in Kiev to suggest the appropriate resilience scenarios to decision makers. At this step of research, we intend to process a large number of Romanian text data that is available on the Internet and is stored in an unstructured manner.

In the following sections, we will describe methods of texts sources processing that permit extracting markers of social disasters. To achieve this, we will gather an amount of data, with the intention to generate a lexicon afterwards, which comprises relevant words with reference to technological, social or natural disasters. Subsequently, based on these markers, we will form a lexicon that serves as basis for our future system of identification and classification of texts from the Internet.

The article consists of several sections. Section 2 introduces several related works that describe the state of the art in our research field.

Section 3 describes the general approach, emphasizing two processing stages. Section 4 presents the collection of articles obtained manually. Section 5 describes the stages of lexicon of markers creation. Sections 6 and 7 are concerned with automation of texts collection and streamline of their processing. The article ends with conclusions and some directions for future works.

2 State of the Art

There are many papers that analyse the topic of social disasters warning and decision making. Social media is considered to be a quick information propagation tool for being informed about recent human kind catastrophes [2]. That is why it is frequently used for disaster monitoring and mitigation [3]. The social networks give the possibility to share the information concerning the damage of disaster [4]. Another idea is the context-aware social networking module for interaction [5], which offers the possibility for posting and locates the place of a disaster in the social network [6]. Moreover there are attempts to elaborate applications for smart phones that would be capable of the disaster response [7].

However, the proposed solutions should also take into account the regional constraints. In [3, 8], the specifics of the problem regarding Romania, Ukraine and the Republic of Moldova is described. We can mention especially the paper [3], where a basis of Romanian Controlled Vocabulary is established. This resource was elaborated being based on a number of professional controlled vocabularies, e.g. those of The International Press Telecommunications Council [9], specialized authorities from Romania, USA and authors experience.

Another example of a professional vocabulary is presented in [10]. It consists of sub vocabularies divided in 17 categories. Every category has several authors representing the authorities that are responsible for the terminology.

Republic of Moldova has also such a Service of Civil Protection and Exceptional Situation classifier [11]. It consists of situations descriptions that are possible in the Republic of Moldova. For our research

this resource serves as a base for social disaster classification.

On the other hand the problem of continuous completion of the Controlled Vocabulary is a permanent task. The keywords that should be included in this vocabulary are taken from social networks, blogs, news online articles etc [12], disaster related keywords being of big importance for emergency system [4].

3 Methodology

One of the first steps of the project is on-going monitoring of the information networks or news sites or social networks used with the aim of finding texts containing any signals about something that has occurred or is about to occur somewhere.

In selecting the relevant information, we can point out relatively distinct approaches for different sources of information: news sites and social networks. In both cases, processing algorithms will take into account the date of publication, because we need to operate with fresh data. Also, it is necessary to exclude various promotional posts and other information having a character of noise, in relation to our topics [13].

At the first glance, it would seem that processing of social networks has an additional key that would allow us to select the posts easily, related to a particular topic – hashtags. They can assist in following chain of posts concerning the given topic. But, as it was mentioned in [14], it is impossible to establish a priori and form a controlled vocabulary of these hashtags, because they operate with a specific lexicon, often unpredictable. Many times hashtags didnt contain any words related directly to disaster subjects, labelling the corresponding event by some toponyms, expressions or some other proper nouns. As examples can serve the well-known hashtag #jesuischarlie or #colectiv related to the fire in a night club in Bucharest, which took dozens of young lives. Therefore we will treat social networks in a manner different from news sites. In this case the controlled vocabulary is formed from two parts: a static one, consisting of keywords selected apriori from different sources and a dynamic part, which comprises relevant hashtags.

The method of their extraction will be described below. To create the above-described tools, we will accomplish the following steps:

Stage I. Preparatory phase.

- Manual Romanian texts collection and their (manual) classification.
- Elaboration of a lexicon of markers based on these texts and other available sources.
- Automated enriching of the vocabulary by flexing and derivation.
- Creating the database with classified texts.

Stage II. Processing phase.

- Automated news texts selection.
- Texts filtering according to lexicon of markers.
- Texts analysis.

In the following sections we will present the approaches that we consider to accomplish these steps.

4 Online News Articles

We used a number of sites from the Republic of Moldova and Romania and managed to collect 616 relevant texts in Romanian. Collected texts were pre-processed and manually classified. 616 news articles were divided into 10 categories of disasters which are present in the Service of Civil Protection and Exceptional Situation classifier: railway, air and cars accidents, fire, earthquake, hurricanes, radioactive sub-stances, attacks, flood and diseases. In case we have a text that can be considered as part of two or more categories, it is assigned to those categories. These 10 categories covering 84% of cases are present in the classifier [11]. The other 16% of the situations are not so frequent, therefore we omitted them. These are related to mass loss of

wildlife, vegetation destruction on a vast territory, considerable change of atmosphere transparency, etc.

They contain 142840 words, from several news sites (see details in Table 1). All texts were lemmatized and annotated with the part of speech tagger. This fact will help us in the identification of those words that refer to social disasters. The next step was to attach, if necessary, a part of sentence, in order to avoid ambiguities and classify them. Finally, we established 10 groups of texts that refer to a specific social situation, with its own set of lexical markers.

Table 1. Statistics on collected articles

Nr.	Category name	Number of articles	Number of words	Article average nr. of words
1	Hurricanes	5	3380	676
2	Earthquake	6	2412	402
3	Radioactive contamination	10	2406	241
4	Diseases	12	5060	422
5	Railway accidents	40	10784	270
6	Air accidents	100	25675	257
7	Cars accidents	100	18005	180
8	Attacks	100	27236	272
9	Flood	103	23922	232
10	Fire	140	24960	178
	Total	616	143840	233

When processing this collection we also applied the disambiguation methods. This is necessary because the words can have multiple mean-

ings occurring in different contexts. For example, the word "bomb" can have a meaning "an explosive weapon detonated by impact" or the meaning of "something very cool/good". Another example would be the word "incendiary", which means both something causing (or designed to cause) fires or something arousing to action (for example, an incendiary speech).

5 Lexicon of Markers

As it was written above, the obtained collection of annotated texts serves as a source for lexicon of markers. Those words that correspond to social disaster topic were manually selected and added to this linguistic resource. In our research, as it was written in [14], we also use some other sources: the site of Service of Civil Protection and Exceptional Situation and Romanian Controlled Vocabulary, developed in [4, 15].

These three sources constitute the main part of lexicon of markers. At the moment we have a lexicon of markers containing more than 350 lemma words. There is a number of samples from this collection: accidenta (injure), alarma (alarm), alerta (alert), avaria (failure), bombarda (bomb), deraia (derailing), detona (detonate), distruge (destroy), evacua (evacuate), exploda (explode), inunda (flooding), nenoroci (perish), ucide (kill), vătăma (hurt), pustii (devastate).

This lexicon of markers needs to be permanently populated by other keywords. The usual way is to increase the number of news articles and to select new words. On the other way we may use the internal linguistic mechanisms to increase the number of words from the lexicon starting with the existent set. That is why a useful component of the system would consist in automatic completion of the lexicon of markers.

The tools we use for text monitoring and analysis operate with word stems. As the Romanian language belongs to the class of inflectional ones, the process of word forming or derivation of a number of vowel or consonant alternations may occur, generating new stems. For example, there are three different stems for Romanian verb "a dărăpăna" (to run-down): dărapăn, dărăpăn, dărapen.

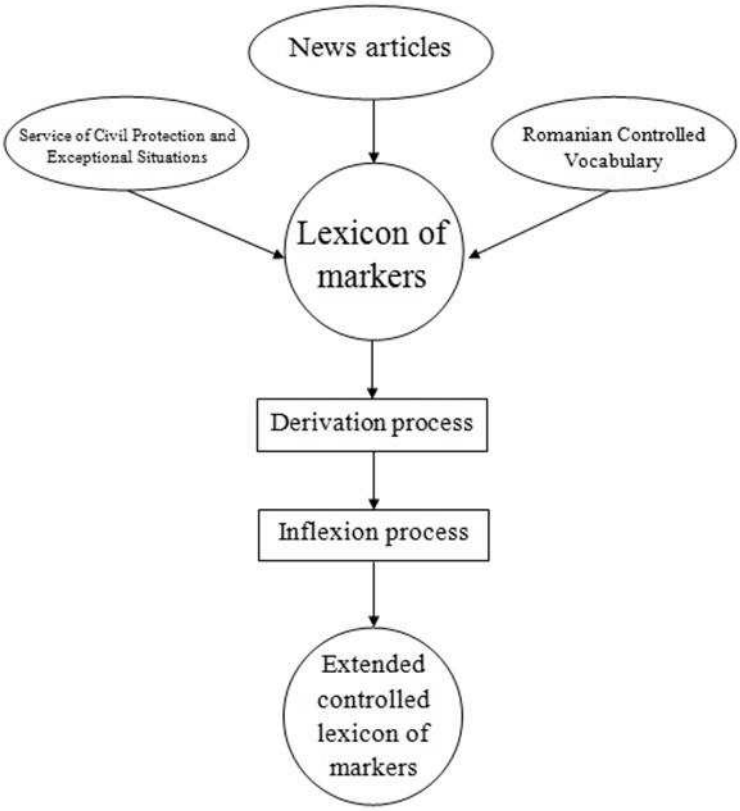

Figure 1. The process of extended controlled lexicon of marker creation

Therefore, for each word it is necessary to have all the possible stems. We use in-house elaborated tool to inflect the selected keywords (it has a general purpose, also applicable in our case). The tool is based on grammar rules with scattered contexts, and word-forms generation is reduced to the corresponding grammar rules interpretation [16]. The inflexion is based on the knowledge about the morphological group of a given word. An algorithm for calculating morphological group of an arbitrary word was developed [17]. Obviously, for each ending we can establish a correspondence with morphological characteristics of the word-form, thus obtaining a possibility of morphological annotation. If in the case of inflection, this does not change the meaning of the obtained words, and the problem is solved automatically [18], in the case of derivation the process of automation is more complicated. Our goal consists in realizing automatic derivation without use of semantic elements in the process of derivation. The only information was concerning the character representation, and in some cases, the part of the speech.

As we described in [19] there are four algorithms: affix substitution, derivatives projection, formal derivation rules and derivational constraints. A few affixes form the overwhelming majority of derivatives: 12 of 41 prefixes formed 88.2% of all derivatives with prefixes, analogously, 52 of the 420 suffixes formed 87.7% of all derivatives with suffixes.

Even if we apply these four algorithms a step of validation is needed. Our approach was based on the Internet filtration and manual validation of the generated words [9]. The method shows that we can increase the vocabulary by approximately 15%, with the accuracy of 89%.

The processes described above refer to the static part of the lexicon. In the case of the dynamic one its completion is performed at processing phase, when tweets containing hashtags are analysed. If a tweet is identified as relevant to disaster topics, its hashtags $H = \{h_1, h_2, ..., h_n\}$ are extracted. It is necessary to determine which of them belongs to the field of interest. For this purpose, all of them are initially followed during m subsequent steps and corresponding tweets are analysed. In case if they refer to the same disasters topic, the hashtags identical to

those contained in the set H are extracted and included in the dynamic part of the controlled vocabulary.

6 Automated Texts Collection

For the processing phase a Crawler-based [20] application service has been elaborated. It inspects various news websites, downloads and extracts the text of this news, and stores it in the database. Crawler is written using Node.js and Request library. Since each site is unique by its structure, plug-ins were written for each of them, taking into account its specific design. Fig. 2 presents the application interface with the result of extracting the text from ProTV news site.

Articles are extracted as follows: RSS-feed is downloaded, and then news is filtered in accordance with the lexicon of markers. Our approach considers the lexicon markers as classifier attributes for the process of classification in those 10 categories. The process of classification is done with j48tree classifier implemented in Weka programme. This approach gives us approximately 76% of accuracy.

After the filtration the full text of news is extracted because RSS-feed is just a part of the article. The process of text extraction is the following: the corresponding page on the website ProTV in HTML format is downloaded, the page with pure text is extracted, then it is cleansed from the rest of HTML tags that remain and only the final result is stored in the database.

7 How to Streamline the Process

In order to automate the process of text collecting referring to disasters, we started with Natural kit for NodeJS. It consists of different natural language tools, including two classifiers, Naive Bayes and logistic regression. Using the established 10 categories we populated a database with collected texts, presenting them in corresponding format. The idea is to have an amount of texts classified in different topics.

The given text is analysed by comparing with the classified texts

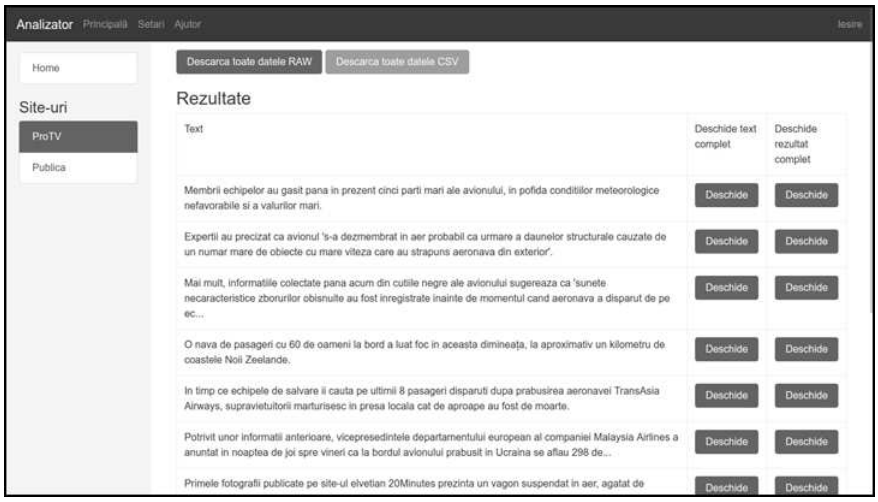

Figure 2. The application interface

from the database, formed by the collected texts. The tool gives a similarity score to a certain text which can offer a statistic idea of how close the analysed text is to a selected topic.

The below example presents the results of processing of the following text: Experţii au precizat că avionul 's-a dezmembrat in aer probabil ca urmare a daunelor structurale cauzate de un număr mare de obiecte cu mare viteză, care au străpuns aeronava din exterior'. (in engl. *Experts have said that the plane 'was dismembered in the air probably due to structural damage caused by a large number of objects with high-speed that have penetrated the aircraft from the outside'.*)

{`label:'Air accidents',value:9.46168293230331e-40`}

{`label:'Railway accidents',value:1.671323536752323e-40`}

{`label:'Radioactive contamination', value:9.51103968862598e-45`}

{`label:'Fire',value:5.94988550817385e-45`}

{`label:'Cars accidents',value:1.232757059054971e-45`}

```
  {label:'Flood',value:6.389181886502171e-49}
node analyze.js 80,83s user 0,04s system 100%
cpu 1:20,87 total
```

One can see that the highest similarity score is achieved in the first case related to "Air accidents", which corresponds to the meaning of the processed text.

One of the observations is that if the database grows by n, then the processing time is growing by $2n$. At the moment the processing time is quite high, e.g. the processing time in the example above is more than one minute. The research direction must be taken to streamline the process. Our goal is to increase the processing speed of texts. One approach is to optimize the information in the database.

We performed the following experiments on a sub-collection of texts (45 news articles, consisting of 11257 words, referring to railway, air and car accidents). The same procedure was applied: annotation at sentence and word levels, providing morphological information using UAIC Romanian Part of Speech Tagger [11]. Based on the obtained results, we got 2659 unique lemmas. In addition, extracting only those which have the frequency more than one, and part of speech noun, verb, adjective and adverb, we obtained 1093 different lemmas. So, the procedure showed how to reduce the number of susceptible words for markers and to optimize the processing time.

On the other hand, in order to reduce processing time, changes were made on the contents of the database, excluding from the collected texts those words or even sentences that are not directly related to the subject of disaster. For example, the following text: "Two persons were killed and seven others injured on Monday evening in a bus explosion in the Armenian capital, Yerevan, announced the Ministry for Emergency Situations of the Caucasian Republic, AFP informs" can be reduced to a short form, namely: "Two persons were killed and seven others injured in a bus explosion". The remainder is sufficient to serve as a pattern for analysis and classification of new extracted texts, but the processing time will be significantly decreased. These minimizing of database records must be made with great accuracy, not to lose substantial information. Obviously, the cuts are operated in patterns

only, not in the news, where, for example, place of the event and source of information can be important for mitigation scenarios.

8 Conclusions and Further Work

Our experience has shown that the proposed tool provides acceptable results. In order to obtain a better classification it is necessary to increase the number of collected texts, especially those related to the topics of hurricanes, earthquake, radioactive contamination and diseases. Despite the optimization measures, the processing time tends to increase with the expansion of the database and we decided to develop a distributed processing algorithm using the 64 node cluster from the Institute of Mathematics and Computer Science.

References

[1] NATO Science for Peace and Security Program. [Online]. Available: http://science.iasa.kpi.ua/sps.

[2] N. O. Hodas, G. V. Steeg, S. Chikkagoudar, J. Harrison, E. Bell, C. D. Corley, "Disentangling the Lexicons of Disaster Response in Twitter," in *WWW 2015 Companion*, Florence, Italy, May 1822, 2015, pp. 1201–1204.

[3] H. N. Teodorescu, "Using analytics and social media for monitoring and mitigation of social disasters," *Procedia Engineering*, vol. 107C, pp. 325–334, 2015.

[4] A. S. Gowri, R. Kavitha, "Tweet Alert: Effective Utilization of Social Networks for Emergency Alert and Disaster Management System," *International Research Journal of Engineering and Technology*, vol. 2, no. 8, 2015, pp. 1065–1070.

[5] R. Simionescu, "Hybrid POS Tagger," in *Proceedings of Language Resources and Tools with Industrial Applications Workshop* (Eurolan 2011 Summer School), Cluj-Napoca, Romania, 2011, pp. 21–28.

[6] R. Rana, I. Kristiansson, I. Hallberg and K. Synnes, "An Architecture for Mobile Social Networking Applications," in *First International Conference on Computational Intelligence Communication Systems and Networks*, CICSYN '09, 2009, pp. 241–246.

[7] B. Brownlee and Y. Liang, *Mobile Ad Hoc Networks: An Evaluation of Smart phone Technologies,* Kingston (ONTARIO), Canada: Royal Military College of Canada, 2011, 40 p.

[8] N. D. Pankratova, P. I. Bidyuk, Y. M. Selin, I. O. Savchenko, L. Y. Malafeeva, M. P. Makukha and V. V. Savastiyanov, "Foresight and Forecast for Prevention, Mitigation and Recovering after Social, Technical and Environmental Disasters," in *Improving Disaster Resilience and Mitigation – IT Means and Tools*, Springer, 2014, pp. 119–134.

[9] J. Hjelm, *Why IPTV: Interactivity, Technologies, Services,* John Wiley & Sons, 2008, 370 p.

[10] U.S. Department of Health & Human Services. Disaster Information Management Research Center. Disaster Glossaries, [Online]. Available: https://disaster.nlm.nih.gov/dimrc/glossaries.html

[11] Service of Civil Protection and Exceptional Situation classifier, [Online]. Available: http://www.dse.md/ro/clasificator

[12] N. D. Pankratova and V.O. Dozirtsiv, "Application of methods for text analysis of the emotional tone to identify social disasters," in *System analysis and information technology: 18-th International conference SAIT 2016*, Kyiv, Ukraine, May 30 – June 2, 2016, pp. 38.

[13] C. Bolea, "Vocabulary, Synonyms and Sentiments of Hazard-related Posts on Social Networks. An analysis for Romanian messages," in *Proceedings of IEEE Conference SPED 2015*, Bucharest, Oct. 2015, pp. 48–53.

[14] M. Petic, S. Cojocaru and V. Gîsca, "Exploring list of markers in automatic unstructured text data processing," in *Proceedings of the 11th International Conference Linguistic Resources and Tools for Processing the Romanian Language*, Iaşi, Romania, 26-27 November 2015, pp. 125–136.

[15] H. N. Teodorescu - private mail

[16] A. Alhazov, E. Boian and S. Cojocaru, "Modelling Inflections in Romanian Language by P Systems with String Replication," in *Proceedings of the 10th Workshop on Membrane Computing*, WMC10, Curtea de Arges, Romania, August 24 27, 2009, pp. 116 –128.

[17] S. Cojocaru and E. Boian, "Determination of infexional group using P systems," *Computer Science Journal of Moldova*, vol. 18, no. 1, pp. 70–81, 2010.

[18] M. Petic and S. Cojocaru, "Vocabulary enriching for text analysis," in *System analysis and information technology: 17-th International conference SAIT 2015*, Kyiv, Ukraine, June 2225, 2015. pp. 37–38.

[19] M. Petic, V. Gîsca and O. Palade, "Multilingual mechanisms in computational derivational morphology," in *Proceedings of Workshop on Language Resources and Tools with Industrial Applications LRTIA-2011*, Cluj-Napoca, Romania, pp. 29–38.

[20] M. Najork and J. L. Wiener, "Breadth-first crawling yields high-quality pages," in *Proceedings of the Tenth Conference on World Wide Web*, Hong Kong, May 2001, Elsevier Science, pp. 114–118.

Svetlana Cojocaru, Mircea Petic,
Grigore Horoş

Received May 15, 2016

Institute of Mathematics and Computer Science
Address: 5, Academiei street, Chisinau, Republic of Moldova, MD 2028
Phone: (373 22) 72-59-82
E–mail: svetlana.cojocaru@math.md, mircea.petic@math.md,
grigori.horos@math.md

Computer Science Journal of Moldova, vol.**24**, no.2(71), 2016

Feasible approach to modeling of ecological component of the sustainable development paradigm

T. V. Shulkevich, Y. M. Selin

Abstract

The paper contains the description of the feasible approach to modeling and forecasting of ecological processes that are the component of the sustainable development paradigm. The suggested mathematical apparatus is based on the statistical methods and comprises hidden Markov models with the linguistic modeling.

Keywords: modeling, forecasting, ecological processes, mathematical apparatus, hidden Markov models, linguistic modeling.

According to the definition suggested by the Centre "For our common future" ("For our common future"), that has been working in Geneva since 1988, and the United Nations Conference in Rio de Janeiro (1992), the sustainable development is understood as the kind of development that "satisfies the current needs, but does not compromise the abilities of the future generations to satisfy their own needs", or that provides "the high quality of environment and the healthy economy for all nations". Therefore, the problem of the sustainable development is the problem of the mankind salvation from the consequences of its own activity that, by the end of the 20th century, have become critical and manifested themselves in desertification, pollution of the atmosphere, oceans and soil, rapid population growth, poverty, starvation, dangerous diseases, etc. The issues of scientific knowledge and education, industry and innovation technologies, ecological, social and medical problems, international relations and political events and many other phenomena of modern life have intertwisted into the single interrelated unit. The consciousness of the situation and the search of the

Figure 1. The triune concept of the sustainable development

reasonable solutions to the problem have become an urgent matter. It requires focused attention not only on a global, but also on a national, regional and local basis [1].

Usually, the sustainable development paradigm is presented in the following way (Fig. 1).

The growing rate of exploitation of natural resources, economic decline of the economies in transition raise up the risk of the technogenic disasters, keep the countries from spending considerable efforts and resources necessary for the implementation of measures to reduce the environmental impact. To accomplish the purpose of the sustainable development the specific mechanisms for its implementation are required. The development and application of the "environmental assessment" system for the assessment of the transition to the sustainable development is of considerable significance in the world practice. The most important are the two of its modes (local and strategic) that nowadays are the well-established mechanisms of revelation and prevention of the environmental consequences of the human induced activity (refer

to Fig. 2) [2].

However, for better understanding of the ecological component of the sustainable development paradigm, it is necessary to analyze some natural phenomena and biospheric processes and their interaction in the "man-environment" system. We will pay careful attention to environmentally hazardous processes.

Environmentally hazardous processes are generally understood as exogenous and endogenous (anthropogenic), short-term and long-term impacts on the ecosystem as a whole or on its particular element, leading to the violation or problems of its performance that in its turn disrupts the man-environment balance.

1. Endogenous (anthropogenic) impacts are caused by anthropogenic activity (pollution by harmful organic and non-organic substances of all kinds of the environment: the atmosphere, surface waters, water-saturated underground layers, soil, animate environments; noise and electromagnetic pollution of the atmosphere; light pollution; flooding of soils and slopes; ozone depletion and creating conditions for the greenhouse effect). They may be controlled and, in some cases, reduced or even prevented.

2. Exogenous impacts are the manifestations of the outer space and internal geological processes and changes (solar cycles, hurricanes, tropical cyclones, monsoons, volcanic activity, global warming as a result of the precession of the Earth's axis, and finally, the asteroid hazard). They are purely objective, cannot be controlled, but may be observed and forecasted. Therefore, it is possible to develop a set of measures to minimize the damages or mitigate the consequences.

It becomes clear from the mentioned earlier that the environmentally hazardous processes are of various nature and types. Also, it is evident that "not everything depends on a man".

In Fig. 3 there is a schematic classification of the environmentally hazardous processes [3].

The solar activity (Fig. 4) is a special type of an exogenous impact on the geological and biochemical processes taking place on Earth. The index for the solar activity characteristics, so called Wolf number, Zurich number or the relative sunspot number is calculated by the

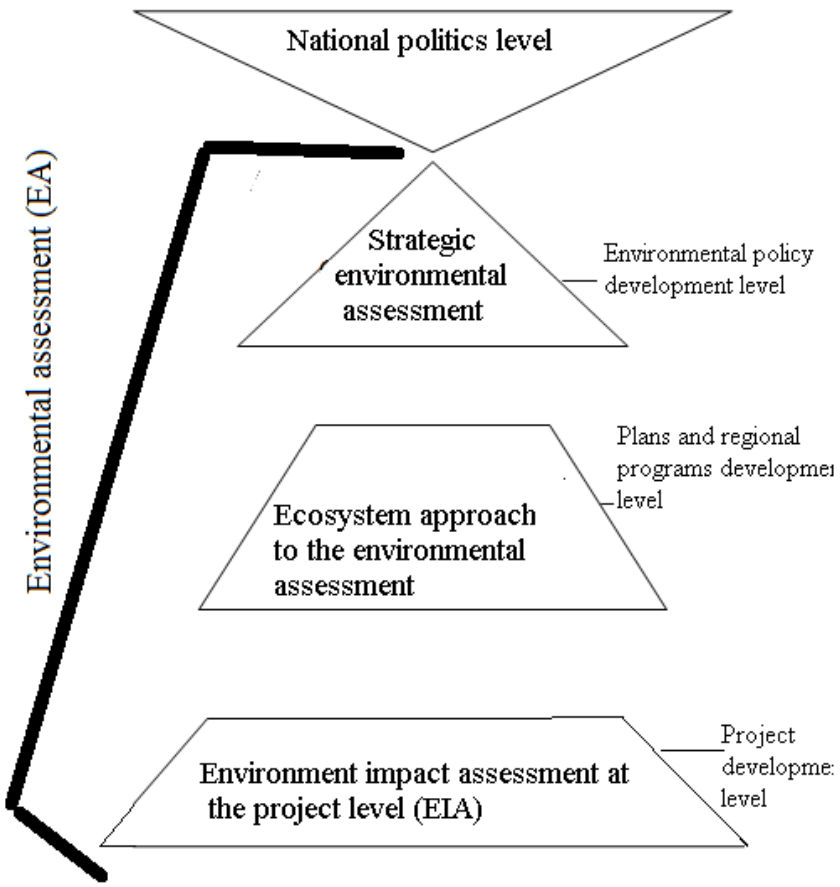

Figure 2. Environmental assessment system (EA)

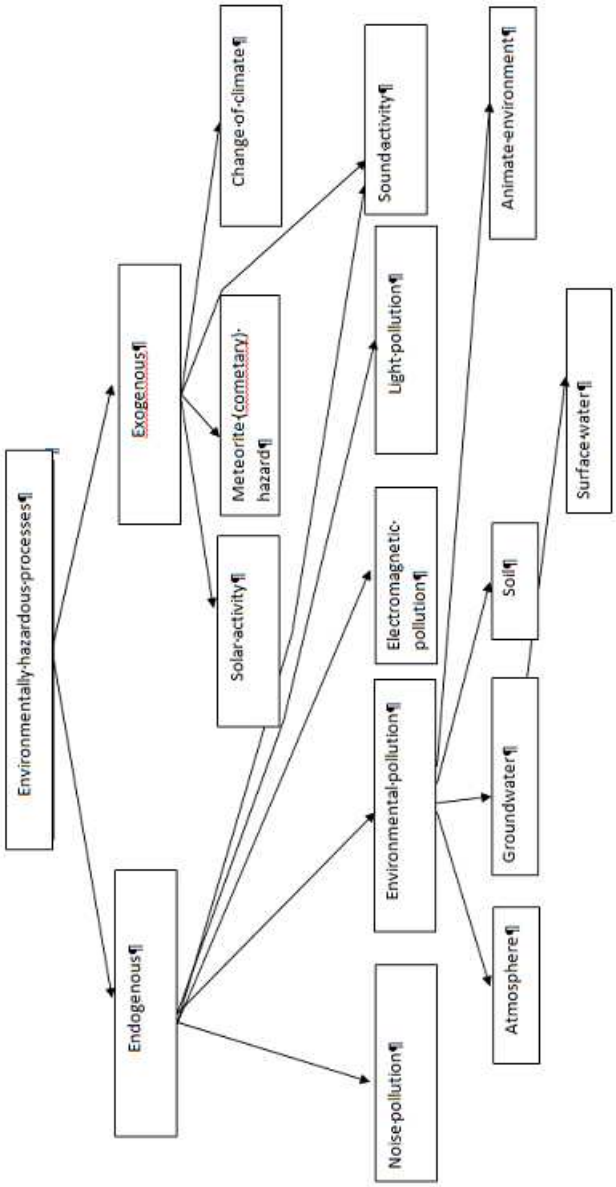

Figure 3. Environmentally hazardous processes classification

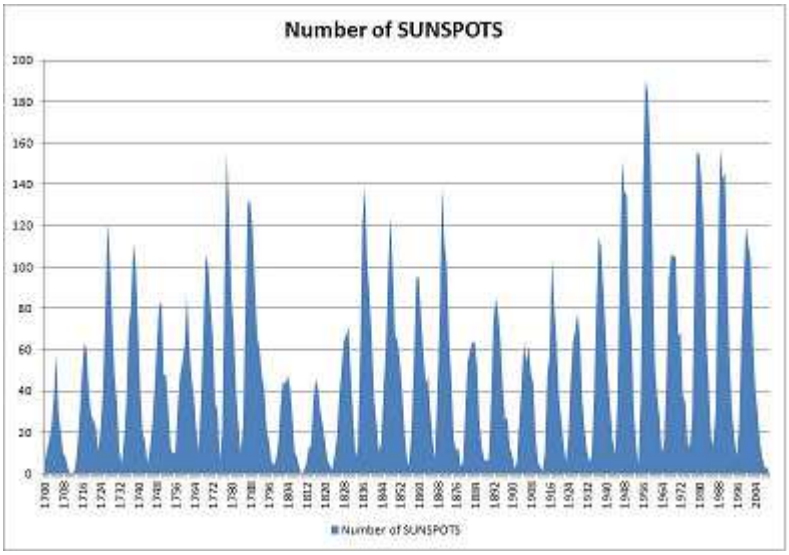

Figure 4. Solar activity cyclical change. Source: US National Centre for Environmental Information site

following formula of $R = k \cdot (f + 10 \cdot g)$, where R is the Wolf number; f is the total number of sunspots on the visible sun hemisphere; g is the number of sunspot groups; k is the multiplier (less than 1) that takes into account the total contribution of observing conditions, the telescope type and that is observed by the standard Zurich numbers.

According to the observation data for the last 304 years, the cycle length in actual practice varies from 8.5 to 14 years between the proximate minimums and from 7.3 to 17 years between the proximate maximums. The connection between the solar cycles and the increased occurrence of land slides has been proved. The most adverse impact is on that of the innovate technology effects, high-frequency operating devices, such as mobile communications, satellites (and also the related navigational safety problems), etc. Solar radiation is the source of the motion in the atmosphere-ocean system. The force, that the solar radiation generates, produces buoyancy – the source of motion

in the atmosphere. It determines the nature of hurricanes, tornadoes, tsunamis and other destructive processes. Also, the solar activity influences the proliferation of some diseases, increases the frequency of traffic accidents, etc.

The climate change caused by the global warming is not only the result of increased greenhouse gases concentrations (especially, carbon dioxide) in the atmosphere, but also the result of the next cycle of the precession of the Earth's axis, thus it has an exogenous cause and it will be intensified in the near future. Although, according to Kh.Y. Shelnhuber [4], Antarctic and Greenland's ice sheet, Sahara desert, Amazonian rainforest, Asian monsoon system, Gulf Stream and six more "weak spots" of Earth influence the climate, and in the case of at least one of them undergoing changes, the ecological disaster (of the continental level) will be inevitable, the risk may be decreased by reducing the emission rate of gases mentioned earlier.

Considering the literature [3], we can define two principal directions and accordingly, two methodological approaches to the mathematical modeling of the dynamics of the environmentally hazardous processes of various nature. The first approach comprises dynamic-computed approaches based on computational methods for solving various kinds of differential equations that describe the basic laws of physics, and also atmospheric and hydrodynamic processes. They are focused on solving the following basic problems of the most important spatiotemporal patterns of the current natural processes:

- Exposure of the current spatiotemporal interrelations between various atmospheric processes in the observation dynamics;
- Natural processes modeling for the forecasting of their development dynamics.

The second approach, comprising empirical dynamic-statistical approaches based on the use of the long-term field measurement statistics, belongs to the international system for the analysis and forecasting of the ecosystem components. They are focused on the exposure of the basic spatiotemporal patterns typical of the atmospheric processes over

decades. The main purpose of these approaches is, in fact, the establishment of deep spatiotemporal correlations between various natural processes based on the long-term statistics. Depending on the purposes of the study it is necessary to perform the development of the mathematical apparatus for the analysis of the dynamics of the environmentally hazardous processes based on either dynamic-computed or dynamic-statistical approaches, taking into account the specific peculiarities and properties of the processes.

Besides, there is the third type of processes that cannot be modeled with the help of dynamic-computed methods, and due to the absence of a peculiar particular periodicity (daily, monthly, annual or another permanent periodicity) they are difficult to describe using empirical-statistical methods. Therefore, the problem of the development of the process analysis and the development of the forecasting methods of such processes for the information support of the environmental situation control and monitoring system is relevant.

To sum up, the conducted analysis allows for the conclusion that the environmentally hazardous processes are characterized by the complex interrelations, interdependencies and interactions of various factors and causes. They have the following characteristic properties and peculiarities:

- The heterogeneousness and diversity of causes and factors and an activity that causes them;
- The spatial distribution of the triggering events, temporal and spatial uncertainty of the growth dynamics and their impact on the eco-surroundings;
- The nonstationarity of properties and ambiguity of their characteristics.

These properties and peculiarities determine the practical relevancy of studying of all the variety of characteristics, interrelations, interactions, interdependencies of the diverse factors and causes of the environmentally hazardous processes on the basis of the single systematic approach from the perspective of achieving the globally defined objective of the environmental situation management – early prevention and

(or) minimization of the adverse effects of their action. However, the analysis shows that nowadays the various types of natural and technogenic environmental processes, their causes, progress, consequences and the sphere of the action are studied separately, without regard to interrelations, interdependencies, and interactions.

Such an approach does not consider some factors of primary importance that influence the active processes, their adverse impact level, the possibility and effectiveness of its prevention.

Considering all of the above-mentioned, we propose the feasible mathematical apparatus that may be used for the forecasting of the environmental processes of all three types [5].

Hidden Markov models (HMM). The following diagram (Fig. 5) shows the general structure of HMM. Ellipses are the variables with the random value. Accidental variable $x(t)$ corresponds to the value of the hidden variable at time t. Accidental variable $y(t)$ is the value of the variable under an observation at time t. The arrows on the diagram stand for the conditional dependences. As it can be seen from the diagram, the value of the hidden variable $x(t)$ (at time t) depends only on the value of the hidden variable $x(t-1)$ (at time $t-1$). It is called the Markov property. However, at the same time the value of the variable $y(t)$ under an observation depends only on the value of the hidden variable $x(t)$ (at time t).

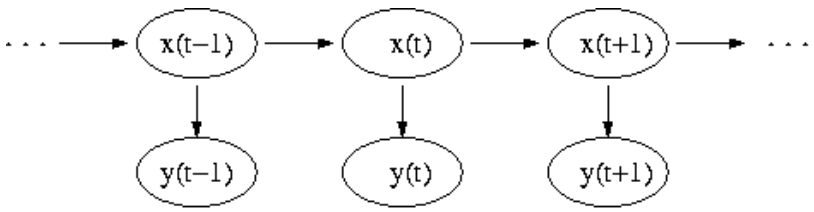

Figure 5. Solar activity cyclical change

The probability of observing the sequence $Y = y(0), y(1), ..., y(L-1)$ of the length L is $P(Y) = \sum_X P(Y/X) \cdot P(X)$.

Here, the sum runs over all possible sequences of hidden points $X = x(0), x(1), , x(L-1)$.

The basic Markov models may be described through the following variables: N is the number of conditions; T is the number of observations; $\theta_{i=1,\ldots,N}$ is the parameter for the observation of the relations between conditions; $\phi_{i=1,\ldots,N;j=1,\ldots,N}$ is the probability of transfer from condition i to condition j; $\phi_{i=1,\ldots,N}$ is N-dimensional vector consisting of j vectors $\phi_{i=1,\ldots,N;j=1,\ldots,N}$; $x_{t=1,\ldots,T}$ is the condition of the observation in a time t; $y_{t=1,\ldots,T}$ is the result of the observation in a time t; $F\left(y/\theta\right)$ is the probability distribution function of observations parameterized by θ.

The result of the environmental process observation with the help of HMM is the sequence of conditions the system undergoes during the time of observation.

Markov model of weather with three conditions will be used as an example. We will consider that the observations are made daily (for example, at noon), the weather may have one of the following conditions:

- S1 – rain (snow);
- S2 – cloudy;
- S3 – clear.

The weather on the day t is described by one of the abovementioned conditions, and the transition-probability matrix has the following form:

$$A = \left(\ a_{ij} \ \right) = \left(\begin{array}{c} 0,4|0,3|0,3 \\ 0,2|0,6|0,2 \\ 0,1|0,1|0,8 \end{array} \right). \tag{1}$$

Thus, having the process model in the form of the probability matrix and the start state probability matrix, we can calculate any condition sequence probability, that is any observation probability.

Linguistic modeling. According to the specific rules the numeric values are replaced with the symbols. The forecasting is made through the search of the symbol strings and their correlation with the strings

from the observation database. Obtained symbols make up the V alphabet from which, in its turn, the words are formed.

Herein, the grammar (or formal grammar) is understood as the way of description of a formal language, that is the discrimination of a certain subset from the whole set of words of a certain finite alphabet.

The forecasting is performed with the help of a stochastic context-free grammar. Stochastic context-free grammar is the context-free grammar in which each deduction rule has a corresponding probability.

Context-free grammar G is the tuple (N, T, P, S): $S \in N$; N and T are the bounded disjoined sets; P is a finite subset $N \times (N \bigcup T)$.

Herein, the following names are used: N is the nonterminal set, T is the terminal set, P is the deduction rules set. Rules $(\alpha, \beta) \in P$ are specified as $\alpha \to \beta$. The left part of the deductive rule must contain one variable (nonterminal symbol). Formally, $\alpha \in N$, $\beta \in (N \cup T) \star \alpha \in N$, $\beta \in (N \cup T)^{\star}$, $\mid \beta \mid \geq 1$ should be realized.

In stochastic context-free grammars the deduction rules correlate with the probability of use:

$$\rho : \mathbf{p} \to \mathbf{R}, \tag{2}$$

where

$$\sum_{\mathbf{p_r} \in \mathbf{P_r}} \rho\left(\mathbf{p_r}\right) = \mathbf{1}. \tag{3}$$

It is necessary to combine both methods to improve the accuracy of the forecast.

Conclusion. Thus, we proposed the mathematical apparatus that might be used for the modeling and forecasting of the environmental processes of the sustainable development paradigm. The common problem of the statistical methods is the lack of the historical information. However, the mankind is engaged in the monitoring of the environment, observation of the weather and natural processes throughout the life. Therefore, the use of the statistical methods in particular is fully justified.

References

[1] United Nations, *Agenda 21-United Nations Environment Programme*, New York, United States: United Nations, April 23, 1993, 300 p. ISBN: 978-92-1-100509-7.

[2] S. V. Patoka, E. V. Khlobistov, "The theoretical basis of environmenttal safety research development and deployment of industrial capacity," in *Ekonomika promyslovosti Ukrainy* (Zb. nauk. prac.), Kiev, Ukraine: RVPS Ukrainy NAN Ukrainy, 2001, pp. 29–37. (in Ukrainian)

[3] Y. N. Selin, "System analysis of ecologically dangerous processes of diferent nature," *System Research and Information Technologies*, no. 2, pp. 22–32, 2007. (in Ukrainian)

[4] S. Rahmstorf, H. J. Schellnhuber, *Der Klimawandel. Diagonose, Prognose, Therapie*, Seoul, Korea: Dosol Publishing Co., 2007, 144 p. ISBN: 3-406-50866-9. (in German)

[5] I. V. Baklan, Y. N. Selin, T. V. Shulkevich, "Mathematical models predict time series of diferent nature," *Vestn. Khersonskogo nacionalnogo universiteta*, vol. 3(50), pp. 213–218, 2014. (in Ukrainian)

Tatiana Shulkevich, Yurii N. Selin

Received April 2, 2016

Tatiana Shulkevich
National Technical University of Ukraine "Kyiv Polytechnic Institute"
Ukraine, Kiev, Peremogy ave., 37
Phone: +380937256753
E–mail: tatyana-victorovna@ukr.net

Yurii N. Selin
Educational-scientific complex "Institute for applied system analysis"
Ukraine, Kiev, Peremogy ave., 37, bild.35
Phone: +380672388649
E–mail: selinyurij@online.ua

Computer Science Journal of Moldova, vol.**24**, no.2(71), 2016

Quantitative analysis of the evacuation system by means of Generalized Stochastic Petri nets*

Titchiev Inga

Abstract

The aim of this article is to perform a quantitative analysis of the evacuation system by using Generalized Stochastic Petri nets and capturing all the properties and characteristics related to its dynamics.

Keywords: modeling, Petri nets, properties verification, quantitative analysis

1 Introduction

To check properties of distributed systems various methods can be used. Petri nets is one of the methods which demonstrated good results. For a more accurate modeling it is required to define a configuration of the system, intended to work in a certain context. It should correspond to certain performance restrictions. Performance restrictions aim to ensure functional characteristics in the current context related to response time. In particular, this study is focused on quantitative investigations related to dynamics of the modeled system.

Social disaster can lead to other accidents and catastrophes and it may be necessary to keep human health and in some cases, human life, and for these it will be opportune to evacuate successfully inhabitants.

The formalism of Petri net can be applied in the both theoretical and practical ways. In order to surprise as close as possible modeled real systems, the classical Petri net has been extended with the notion of time [1], [3]. Petri nets are a powerful modeling technique because

* This work was supported by NATO project Ref. Nr. SPS G4877

they can be used to model complex systems and to verify if the modeled systems satisfy some criteria.

In order to perform a case study of extended evacuation system we used analysis modules of PIPE [2] and obtained properties and characteristics of them.

In this way the formal method like Petri nets becomes an important tool for detecting, monitoring, modelling and mitigating social disasters [5], [6] caused by actions of different nature.

The paper is organized as follows. First, existing approaches related to modelling disaster and emergency management activities will be presented. Section 2 deals with the requirements of modeling using Generalized Stochastic Petri nets. In Section 3 the proposed model for the evacuation of people in case of disaster will be introduced, including a case study. In Section 4 quantitative analysis is presented with the obtained results. The paper finalizes with conclusions.

2 Generalized Stochastic Petri Nets

We will use the Generalized Stochastic Petri Nets (GSPN) [4] to perform quantitative analysis, because they can capture aspects of production in time of actions and immediately produce some actions. They are characterized by two types of transitions:

1. *Stochastic transitions*: associated with an exponentially distributed firing delay;
2. *Immediate transitions*: associated with a null firing delay.

Formally, a GSPN can be defined as follows: GSPN $= (P; T; \Pi; I; O; H; M_0; W)$, where

- P is a set of places;
- T is a set of transitions, $P \bigcap T = \emptyset$;
- $I; O; H : T \rightarrow N (N = P \bigcup T)$, are the input, output and inhibition functions;

- $M_0 : P \to N$ is the initial marking;
- $\Pi : T \to N$ is the priority function that associates the lower priorities to timed transitions and higher priorities to immediate transitions.
- $W : T \to R$ is a function that associates a real value to the transitions, $w(t)$ is:
  - a (possibly marking dependent) rate of a negative exponential distribution specifying the firing delay, when transition t is a timed transition (represented by a hollow rectangle).
  - a (possibly marking dependent) firing weight, when transition t is immediate (represented by a filled rectangle).

When a new marking is reached, if only timed transitions are enabled, this marking is called *tangible*; if at least one immediate transition is enabled, the marking is called *vanishing*.

The selection the transition of which will fire is based on the priorities and weights. First, the set of transitions with the highest priority is found and if it contains more than one enabled transition, the selection is based on the rates or weights of the transitions according to the expression:

$$P(t) = \frac{w(t)}{\sum_{t\prime \in E(M)} w(t\prime)}, \tag{1}$$

where $E(M)$ is the set of transitions enabled at marking M, i.e. the set of enabled transitions with the highest priority.

3 Evacuation system

Suppose that there is a building as it is specified in Fig. 1. $M1 - M8$ are the rooms, $M11 - M81$ are the doors. Petri Net of this building is given in Fig. 2.

Rooms are modeled using places $R1 - R8$, doors are modeled using places $D11 - D81$, movements from the rooms to the doors are

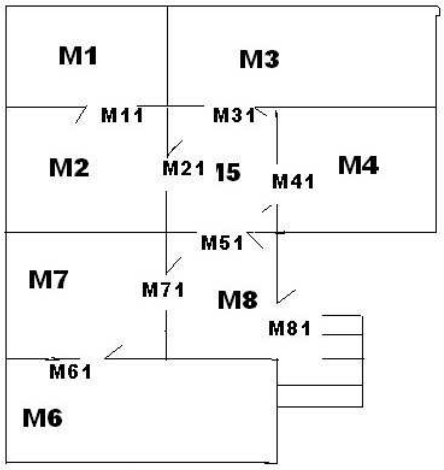

Figure 1. Building plan

modeled by timed transitions $t_0 - t_7$, movements from the doors into the rooms are modeled by immediate transitions $t_8 - t_{14}$. Each inhabitant is modeled by one token, respectively. People are accumulated in the places. The transfer function is used, which takes into account the time spent in the queue (moving time of a human in the room) and the density and flow rate. The initial marking is $M_0 = (1, 1, 0, 0, 0, 0, 1, 0, 3, 0, 2, 0, 1, 0, 0, 0)$.

4 Quantitative analysis

After the simulation we obtained the results from which it is observed the exponential growth of the number of tangible states (Table 1).

The average number of tokens (people) was also obtained (Fig. 3). The number of people from the initial state is constant.

The net is bounded, has a finite set of states (1124 states) which lead to a finite number of steps necessary for evacuation. Also it is safe, transitions do not influence each other, each place works independently

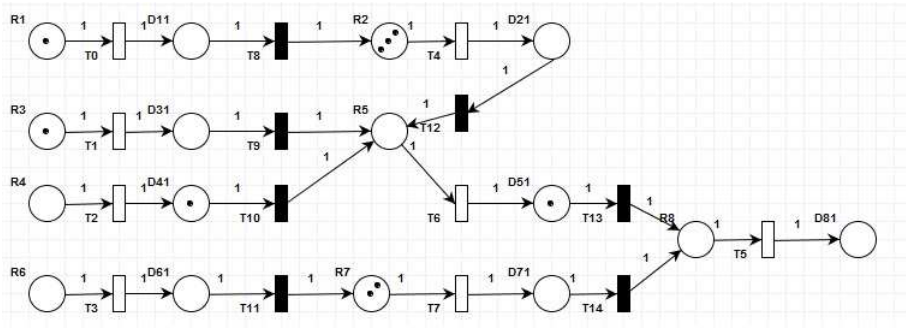

Figure 2. GSPN which models the building from Fig. 1

Table 1. Tangible states

Nr	Number of people in rooms	Number of tangible states	Number of arcs in reachability graph
1	4	241	364
2	9	1124	2389
3	15	1888	3086

Petri net simulation results

Place	Average number of tokens	95% confidence interval (+/-)
R1	0.35	0.2147
R3	0.1	0.21825
R4	0	0
R6	0	0
D11	0.025	0
D31	0.025	0
D41	0	0.024
D61	0	0
R2	0.475	0.93823
R5	2.375	1.35811
R7	0.55	0.45015
D21	0.1	0
D51	0.175	0.024
D71	0.05	0
R8	3.325	1.37582
D81	1.45	1.04339

Figure 3. Average number of tokens in places

Throughput of Timed Transitions

Transition	Throughput
T0	0.28205
T1	0.07692
T2	0.28205
T3	0.28205
T4	0.28205
T5	0.35897
T6	0.35897
T7	0.07692

Figure 4. Throughput of timed transitions

P-Invariants

D11	D21	D31	D41	D51	D61	D71	D81	R1	R2	R3	R4	R5	R6	R7	R8
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

The net is covered by positive P-Invariants, therefore it is bounded.

P-Invariant equations

M(D11) + M(D21) + M(D31) + M(D41) + M(D51) + M(D61) + M(D71) + M(D81) + M(R1) + M(R2) + M(R3) + M(R4) + M(R5) + M(R6) + M(R7) + M(R8) = 9

Figure 5. P invariants

of one another. It is conservative (Fig. 5), the number of people is constant, new people do not appear, all 9 people from the initial state have been accumulated in the last place $D81$.

5 Conclusion

In this study, a method of Generalized Stochastic Petri Nets was proposed for modeling and simulation of system that represents emergency evacuation of people in case of social disaster. This method allows checking such properties as boundedness, conservativeness, deadlock, safeness.

References

[1] W.M.P van der Aalst, "Interval Timed Coloured Petri Nets and their Analysis," in *Application and Theory of Petri Nets* (Lecture Notes in Computer Science, vol. 691), M. Ajmone Marsan, Ed. Berlin: Springer-Verlag, 1993, pp. 453–472.

[2] N. Akharware, "PIPE2: Platform Independent Petri Net Editor," M.S. thesis, Department of Computing, Imperial College of Science, Technology and Medicine (University of London), London, 2005. [Online]. Available:

http://pipe2.sourceforge.net/documents/PIPE2-Report-20050814.pdf.

[3] B. Berthomieu and M. Diaz, "Modelling and verification of time dependent systems using Time Petri Nets," *IEEE Transactions on Software Engineering*, vol 17, no. 3, pp. 259–273, Mar. 1991.

[4] E. Gutuleac, *Performance Evaluation of Computer Systems by Stochastic Petri Nets*, Chisinau, Moldova: Technical Info, 2004, 276 p. (in Romanian)

[5] M. Takashi, N. Yoshifumi, F. Yasuhiro and M. Atsushi, "Development of Tsunami refuge PETRI-NET simulation system utilizable in independence disaster prevention organization, " in *The* 14th *World Conference on Earthquake Engineering* October 12-17, Beijing, China, 2008, Paper ID: 09-02-0018. [Online]. Avalaible: http://www.iitk.ac.in/nicee/wcee/article/14_09-02-0018.PDF

[6] I. Titchiev, "Petri nets to model disaster prevention," in *Proceedings of the Workshop on Foundations of Informatics*, Chisinau, Republic of Moldova, Aug. 2015, pp. 445–449.

Inga Titchiev

Received August 2, 2016

Inga Titchiev
Institute of Mathematics and Computer Science
Academiei street, 5 nr.
Phone: 022-73-80-71
E–mail: inga.titchiev@gmail.com

Computer Science Journal of Moldova, vol.**24**, no.2(71), 2016

The general prioritization framework

Alexey Malishevsky

Abstract

This paper proposes the general prioritization framework for test case prioritization during regression testing. Regression testing (RT) is done to ensure that modifications have not created new faults or that modifications fulfilled their intended purpose by correctly altering software functionality. Being performed multiple times, RT can have a profound effect on the software budget. The test case prioritization orders test cases for execution to reach a certain objective. Usually, such an objective is to detect faults as early as possible during the testing process. Many prioritization techniques have been developed that successfully reach this objective. However, most of these techniques were developed and studied independently from each other despite the fact that they have many similarities. This article presents the framework that allows to represent known prioritization techniques. Thus, it helps to improve existing and devise new techniques. Also, it allows to implement a single tool that emulates any prioritization technique by just setting the correct parameters. The proposed framework includes the combination/condensation (CC) structure and the structure functions including *element combination functions*, *condensation functions*, and a *super-group combination function*. By defining two such structures together with the corresponding structure functions, one for computing award values and one for their update, any known prioritization technique can be expressed. A general prioritization algorithm is presented that can express any known prioritization technique.

Keywords: Prioritization, regression testing, software testing, prioritization framework, test case prioritization.

1 Introduction

Each time a software system is modified and is to be released, it is regression tested. Regression testing (RT) is similar to testing in general: it involves

executing tests and checking the results for correctness. RT, however, is done to ensure that modifications have not created new faults or that modifications fulfilled their intended purpose by correctly altering software functionality.

Being performed multiple times, RT can have a profound effect on the software budget. Because RT itself accounts for a large percentage of software cost [1, 9], even small reductions in RT cost can have a profound effect on the software cost.

If engineers must execute all test cases, which order of test cases should be used? One test order can be better than another under some metric. Test case prioritization orders a test suite to maximize some objective function defined on test orderings. The test case prioritization problem is defined as follows: given a test suite T, the set of permutations PT of T, and a function f from PT to the real numbers; the problem is to find $T' \in PT$ such that $(\forall T'')\ (T'' \in PT)\ (T'' \neq T')\ [f(T') \geq f(T'')]$, where PT is the set of possible prioritizations (orders) of T, and f is an objective function that, applied to any such order, yields an *ordering quality* value for that order.

There are many possible goals for prioritization. For example, testers may wish to increase the coverage of code in the system under test at a faster rate, increase their confidence in the reliability of the system at a faster rate, or increase the rate at which test suites detect faults in that system during regression testing. In the definition of the test case prioritization problem, f represents a quantification of such a goal.

In the literature, many prioritization techniques have been proposed and their effectiveness studied. We will mention just a few of them. Elbaum et al.[6, 7, 8] and Rothermel et al. [15, 16] proposed a set of modification-, coverage-, and fault-exposing-potential-based prioritization techniques. Elbaum et al. [5] incorporated tests costs and fault severities in prioritization. Malishevsky et al. [11] proposed the cost-benefits model for prioritization. Do et al. [4] introduced time constraints into prioritization. Mei et al. [12] applied prioritization to service-oriented business applications. Do et al. [3] explored the usage of mutation faults in prioritization. Bryce et al. [2] utilized prioritization for event-driven software. Raju et al. [14] based prioritization of test cases on four factors such as the rate of fault detection, requirements volatility, fault impact, and implementation complexity. Zhang et al. [18] used integer linear programming for time-aware prioritization. Also, Walcott

et al. studied time-aware prioritization in [17]. Mirarab et al. [13] employed Bayesian Networks for the test case prioritization. Hla et al. [10] used particle swarm optimization methods for prioritization.

Most of the proposed prioritization methods were developed and studied independently from each other despite the fact that they had a lot of similarities. We exploited these similarities among prioritization techniques to develop a unifying prioritization framework. This framework can express every prioritization technique developed so far. Its main benefits include the ability to facilitate creation of new prioritization techniques and their analysis, while providing a standard way of looking at techniques. This framework allows us to implement a general prioritization algorithm whose parameters can instantiate various prioritization techniques. It allows rapid prototyping of and research on a variety of new prioritization techniques with minimal coding, while shortening the time to study them, encouraging experimentation with development of new techniques, and reducing the number of errors that might occur if every technique is to be implemented from scratch.

2 Combination/Condensation Structure

We now formally define the combination/condensation (CC) structure on which our framework is based. We first define an *element* e. An element represents a single piece of data used by a prioritization technique. For example, an element $e \in \mathbb{E}$ can represent coverage information for a given statement s, modification information (number of lines changed) for function f, or the fault-exposing-potential for location l. A single element, however, represents this information for the *whole test suite*; thus, it contains such data for every test from the test suite. We also define the set $\mathbb{E}$ as the set of all elements used in the combination/condensation structure.

We define *sub-element* e^t to be a constituent part of element e corresponding to a given test t. We represent an *element* $e \in \mathbb{E}$ as a tuple $< e^1, e^2, ..., e^{|T|} >$ of size $|T|$, where T is a test suite. For example, if element e represents coverage information for statement s, each sub-element e^t represents coverage information for statement s with respect to test case t.

A *vector* v in our structure is a one dimensional array of elements. More formally, $v =< c_1, c_2, ..., c_{|v|} >$, where $\forall c_l \; 1 \leq l \leq |v| \quad \exists e \in \mathbb{E} \quad c_l \equiv e$.

We define a set of elements that comprise a vector v to be E_v.

We define a *group* G to be a tuple of vectors, $G =< v_1, v_2, ..., v_{|G|} >$.

We define V to be a set of all vectors across all groups. $\bigcup_{v \in V} E_v \subseteq \mathbb{E}$, but this subset may be proper: some elements may not belong to any vector. We define such elements as *free* (e is free iff $\forall v \in V \;\; e \notin E_v$). All vectors that belong to the same group must be *compatible*, defined as having the same number of elements. Vectors across different groups need not be compatible.

For each group G, we define $M_G = (m_{i,j})$ to be the matrix whose columns are the vectors in G, so $m_{i,j}$ represents the i-th component of the j-th vector in G.

Finally, on the top level, a *super-group* SG is defined as a set of groups.

Informally, each element represents a single item of information used by a prioritization technique. It may include function coverage, test costs [5], module criticalities [5], change information [7], or fault-exposing-potential information [6]. Vectors usually represent sets of elements that are treated in the same way, such as coverage, change information, fault-exposing-potential, etc. Groups usually represent sets of vectors used to compute a single component of the final value (produced by the structure).

Next, we define several functions that operate on this CC structure. First, for each group G containing $|G|$ vectors, we define an *element combination function*

$$f^G_{\text{element\_combine}}(x_1, x_2, ..., x_{|G|}, \alpha) \tag{1}$$

This function takes a slice $M^G_{i,1...|G|}$ (an array $< x_1, x_2, ..., x_{|G|} >$ where x_l is the i-th component of the l-th vector in G). This function also has argument α which will be explained later. Each group has its own function $f^G_{\text{element\_combine}}$.

Next, for each group G, we define a *condensation function*

$$f^G_{\text{condensation}}(y_1, y_2, ..., y_k, \alpha), \tag{2}$$

where k is the number of elements in each vector in group G. This function takes the array $< y_1, y_2, ..., y_k >$, where each y_i is the result of applying an *element combine* function to the i-th slice of matrix M_G (defined earlier) ($y_i = f^G_{\text{element\_combine}}(m_{i,1}, m_{i,2}, ..., m_{i,|G|}, \alpha) \quad \forall i, \quad 1 \leq i \leq k$), and an argument α (explained later). Each group has its own *condensation function*.

Finally, we define a *super-group combination function*

$$f_{\text{group\_combine}}(z_1, z_2, ..., z_{|SG|}, \alpha) \tag{3}$$

This function takes an array $< z_1, z_2, ..., z_{|SG|} >$ (z_j corresponds to group $G_j \in SG$) and an argument α (explained later). Each z_j is produced by a *condensation function*: $z_j = f^{G_j}_{\text{condensation}}(y_1, y_2, ..., y_k, \alpha)$.

We call the *element combination function*, *condensation function*, and *super-group combination function*; *structure functions* (SF).

We call this sub-element/element/vector/group/supergroup structure, together with the structure functions, a *combination/condensation structure*.

Let CCF be a particular combination/condensation structure. We define a function $F_{CC}(CCF, \mathbb{E}, \alpha)$ which takes CCF, set $\mathbb{E}$ of elements, and an argument α (explained later), and produces the result of applying the structure functions to the elements of $\mathbb{E}$.

3 Framework and algorithm

In the framework that we now present, we use an iterative approach in which we apply combination/condensation structures to compute award values[1] and to alter elements each time a new test is selected. The main idea is to modify elements $e \in E$, given a newly selected test. This framework uses two CC structures: one, CCF_{award}, for award value computation and another, CCF_{update}, for updating elements from $\mathbb{E}$, where $\mathbb{E}$ is the set of all elements used in the framework.

To compute award values, for each test $t \in T$, we obtain $a_t = F_{CC}(CCF_{\text{award}}, E, t)$. To select the test with the best award value, we compute $t_s = f_{\text{best}}(\vec{a})$. Function f_{best} takes a vector of award values $\vec{a}$ and finds the test (id) with the best award value.

After a new test is selected, the framework algorithm updates all elements $e \in \mathbb{E}$. To compute a new value for an element e, several steps are taken. First, for each test $t_u \in T$ and for each element $e \in \mathbb{E}$, we compute $u_e^{t_u} = F_{CC}(CCF_{\text{update}}, E, < t_s, t_u, e >)$, using CC structure CCF_{update}, where t_s is the test selected in the previous step. Second, we update every element

[1] An award value is the measure of test case's "worth".

Algorithm 1 The prioritization framework algorithm.

```
1:  Initialize elements in 𝔼
2:  List = ϵ
3:  for all t_u ∈ T do
4:      for all e ∈ 𝔼 do
5:          x_e^{t_u} ← f_update(e^{t_u}, F_CC(CCF_update, E, < nil, t_u, e >))
6:      end for
7:  end for
8:  for all t_u ∈ T do
9:      for all e ∈ 𝔼 do
10:         val(e^{t_u}) ← x_e^{t_u}
11:     end for
12: end for
13: loop
14:     for all t ∈ T do
15:         a_t ← F_CC(CCF_award, E, t)
16:     end for
17:     t_s ← f_best(a⃗)
18:     if a_{t_s} = nil then
19:         HALT
20:     end if
21:     Add t_s into List
22:     for all t_u ∈ T do
23:         for all e ∈ 𝔼 do
24:             x_e^{t_u} ← f_update(e^{t_u}, F_CC(CCF_update, E, < t_s, t_u, e >))
25:         end for
26:     end for
27:     for all t_u ∈ T do
28:         for all e ∈ 𝔼 do
29:             val(e^{t_u}) ← x_e^{t_u}
30:         end for
31:     end for
32: end loop
```

$e \in \mathbb{E}$ for every test $t_u \in T$ $val(e^{t_u}) = f_{\text{update}}(e^{t_u}, u_e^{t_u})$. We define $val(x)$ to be the value of x.

The framework is applied as follows: determine the set of elements E in all structures, determine the CC structure CCF_{award} for award computation containing a subset of E, determine CC structure CCF_{update} for update computation containing a subset of E, determine initial values for all elements in $\mathbb{E}$, determine a function for element updating f_{update}, decide on a sorting function f_{best}, and finally, apply the framework algorithm that initializes all elements and computes award values, selects a test, and updates elements until the halting condition is satisfied.

The framework algorithm that implements prioritization techniques is presented as Algorithm 1. Lines 3-7 compute initial values for every element. Lines 8-12 set element values to the values computed in lines 3-7.

Group1		Group2	
V1	V2	V1	V2
covF1	bfiF1	covF1	fepF1
covF2	bfiF2	covF2	fepF2
covF3	bfiF3	covF3	fepF3
covF4	bfiF4	covF4	fepF4
covF5	bfiF5	covF5	fepF5

Figure 1. The structure for prioritization technique *fn-bfi-fep-nofb*.

Lines 13-32 implement the main loop computing the prioritized test case sequence. Lines 14-16 compute test award values. Line 17 finds the test with the highest award value. Lines 18-20 test the halting condition. Line 19 adds the selected test to the ordered sequence. Lines 22-26 compute the new values of elements. Lines 27-31 update the value of every element using values computed in lines 18-20. We define the nil value to be the lowest award value a test can have. During comparisons, a test case with award value nil can be chosen only if there are no tests in T with award values not equal to nil.

4 An example

Now we will demonstrate how to fit one existing prioritization technique into the framework. The *fn-bfi-fep-nofb technique* prioritizes tests in an order of decreasing values of sum of covered fault-exposing-potential[2] of modified functions [7]. Thus, this technique employs a binary fault index[3] and fault-exposing-potential information. There are two vectors in each of the two groups: coverage and binary fault index vectors in the first group, and coverage and fault-exposing-potential vectors in the second group. In each group, corresponding elements of that group's two vectors are multiplied and summed. Then, for each test case, an award value is created as a tuple consisting of two values, one from each group; this is used to order test cases. Award values, being tuples, are compared element-wise: first elements are used for sorting, and, in a case of a tie, second elements are compared. The

[2]The probability that a given test case reveals a fault in a given location, if one exists [7].
[3]Binary fault index is the metric on a code change [7].

combination/condensation structure for the award value computation for this technique is presented in Figure 1 where the *element combination* function is multiplication, the *condensation* function is summation, and the *group combination* function is tuple creation. $CovFi$ is the binary function coverage information for function Fi, $bfiFi$ is the binary fault index for function Fi, and $fepFi$ is the fault exposing potential for function Fi. Each of $covFi$, $bfiFi$, and $fepFi$ is a vector of size $|T|$ whose components correspond to test cases from test suite T. As we can see, these techniques fit easily into the framework.

5 Conclusions

We exploited similarities among prioritization techniques to develop a unifying prioritization framework. This framework can express prioritization techniques developed so far. This framework helps to create new prioritization techniques and analyse them, while providing a standard way of looking at techniques. This framework allows us to implement a general prioritization algorithm whose parameters can instantiate various prioritization techniques. It allows rapid prototyping of techniques and research on a variety of new techniques with minimal coding, shortening study time, encouraging experimentation with development of new techniques, and reducing the number of errors that might occur if a technique is to be implemented from scratch.

References

[1] B. Beizer, *Software Testing Techniques*. Van Nostrand Reinhold, New York, NY, 1990.

[2] R. Bryce, S. Sampath, and A. Memon, “Developing a single model and test prioritization strategies for event-driven software.” *IEEE Trans. on Softw. Eng.*, Vol. 37, No. 1, pp. 48–64, Jan.-Feb. 2011. Available: doi: 10.1109/TSE.2010.12.

[3] H. Do and G. Rothermel, “On the use of mutation faults in empirical assessments of test case prioritization techniques.” *IEEE Trans. Softw. Eng.*, Vol. 32, No. 9, pp. 733–752, Sep. 2006. Available: doi: 10.1109/TSE.2006.92.

[4] H. Do, S. Mirarab, L. Tahvildari, and G. Rothermel, "The effects of time constraints on test case prioritization: A series of controlled experiments." *IEEE Trans. on Softw. Eng.*, Vol. 36, No. 5, pp. 593–617, Sep./Oct. 2010. doi: 10.1109/TSE.2010.58.

[5] S. Elbaum, A. Malishevsky, and G. Rothermel, "Incorporating varying test costs and fault severities into test case prioritization." in *Proc. of the Intern. Conf. on Softw. Eng.*, 2001, pp. 329–338 .

[6] S. Elbaum, A. Malishevsky, and G. Rothermel, "Test case prioritization: A family of empirical studies." *IEEE Trans. of Softw. Eng.*, Vol. 28, No. 2, pp. 159–182, Feb. 2002.

[7] S. Elbaum, P. Kallakuri, A. G. Malishevsky, G. Rothermel, and S. Kanduri, "Understanding the effects of changes on the cost-effectiveness of regression testing techniques." *Journal of Software Testing, Verification, and Reliability*, Vol. 13, No. 2, pp. 65–83, June 2003.

[8] S. Elbaum, G. Rothermel, S. Kanduri, and A. G. Malishevsky, "Selecting a cost-effective test case prioritization technique." *Softw. Quality J.*, Vol. 12, No. 3, pp. 185–210, Sep. 2004. Available: doi: 10.1023/B:SQJO.0000034708.84524.22.

[9] C. Ghezzi, M. Jazayeri, and D. Mandrioli, *Fundamentals of Software Engineerijng*. 1st ed., Upper Saddle River, NJ: Prentice Hall, 1991.

[10] K. Hla, Y. Choi, and J. Park, "Applying particle swarm optimization to prioritizing test cases for embedded real time software retesting." in *IEEE 8th Intern. Conf. on Computer and Inform. Technology Workshops*, pp. 527–532, 2008. Available: doi: 10.1109/CIT.2008.Workshops.104.

[11] A. Malishevsky, G. Rothermel, and S. Elbaum, "Modeling the cost-benefits tradeoffs for regression testing techniques." in *Proceedings of the International Conference on Software Maintenance*, 2002, pp. 204–213.

[12] L. Mei, Z. Zhang, W. K. Chan, and T. H. Tse, "Test case prioritization for regression testing of service-oriented business applications." in *Proc. of the 18th Intern. Conf. on World Wide Web*, WWW '09, 2009 pp. 901–910, New York, NY, USA: ACM. Available: doi: 10.1145/1526709.1526830.

[13] S. Mirarab and L. Tahvildari, "An empirical study on bayesian network-based approach for test case prioritization." in *2008 1st Intern. Conf. on*

Softw. Testing, Verification, and Validation, 2008, pp. 278–287, Available: doi: 10.1109/ICST.2008.57.

[14] S. Raju and G.V. Uma, "An efficient method to achieve effective test case prioritization in regression testing using prioritization factors." *Asian J. of Inform. Tech.*, Vol. 11, No. 5, pp. 169–180, 2012. Available: doi: 10.3923/ajit.2012.169.180.

[15] G. Rothermel, S. Elbaum, A. G. Malishevsky, P. Kallakuri, and X. Qiu, "On test suite composition and cost-effective regression testing." *ACM Trans. Softw. Eng. Methodol.*, Vol. 13, No. 3, pp. 277–331, July 2004. Available: doi: 10.1145/1027092.1027093.

[16] Gregg Rothermel, Sebastian Elbaum, Alexey Malishevsky, Praveen Kallakuri, and Brian Davia, "The impact of test suite granularity on the cost-effectiveness of regression testing." in *Proc. of the 24th Intern. Conf. on Software Engineering*, 2002, pp. 230–240.

[17] Kristen R. Walcott, Mary Lou Soffa, Gregory M. Kapfhammer, and Robert S. Roos, "Time-aware test suite prioritization." in *Proc. of the 2006 Intern. Symp. on Softw. Testing and Analysis*, ISSTA '06, 2006, pp. 1–12, New York, NY, USA: ACM. Available: doi: 10.1145/1146238.1146240.

[18] L. Zhang, S. Hou, C. Guo, T. Xie, and H. Mei, "Time-aware test-case prioritization using integer linear programming." in *Proc. of the 18th Intern. Symp. on Softw. Testing and Analysis*, 2009, pp. 213–224. Available: doi: 10.1145/1572272.1572297.

Alexey Malishevsky

Received March 29, 2016

Institution: ESC "Institute for Applied System Analysis" National Technical University of Ukraine "KPI"
Address: building 35, 37 Prospect Peremohy, 03056, Kyiv, Ukraine
Phone: +380504101177
E–mail: alexeym_s@yahoo.com

Computer Science Journal of Moldova, vol.**24**, no.2(71), 2016

Multilayered Knowledge Base for Triage Task in Mass Casualty Situations

Svetlana Cojocaru, Constantin Gaindric, Iulian Secrieru, Sergiu Puiu, Olga Popcova

Abstract

Use of portable ultrasound becomes common practice in mass casualty situations. The obtained information helps emergency crews to make decisions regarding on-site triage, helping in determination of adequate diagnostics in proper time for saving lives of patients.

In this article a design of multilayered knowledge base in domain of the abdominal region ultrasound examination for the case of mass casualty situations is described. Layers 1-2 correspond to casualty's severity state, and layer 3 – to pathologies when the free fluid is present in the abdominal cavity, that is not the consequence of an abdominal injury.

Keywords: Knowledge base re-engineering, on-site triage, mass casualty, medical ultrasound, hepato-pancreato-biliary region.

1 Introduction

The existence of people and society is increasingly being subjected to serious challenges caused by catastrophes, disasters or natural emergency situations (earthquakes, landslides, floods, etc.), technogeneous, biological, social and premeditated (terrorism) ones.

A *disaster* is a natural or man-made event that suddenly or significantly disrupts normal community function and causes concern for the safety, property and lives of the citizens. Thus, disaster is an event that exceeds the capabilities of the response.

Since 20th century due to technological progress the number of disasters considerably increased, as well as their magnitude. In disaster focus an enormous number of victims died before hospitalization, having injuries compatible with life, because healthcare services did not provide a full qualitative rapid aid.

A *mass casualty incident* is an event that exceeds the health care capabilities of the response, when health care needs additional large resources.

The problem of providing medical aid in the case of a large number of victims was understood in 1881. As a consequence of a fire at the Ring Theater in Vienna more than 400 persons with trauma and burns did not obtained medical aid up in the morning because of lack of overnight medical emergency service.

In case of mass casualty situations, the frequency of which is growing, the number of victims usually exceeds local medical resources. Terrorist attacks of great resonance – attacks on the World Trade Center in New York, bombings in Madrid and London – have resulted in large numbers of victims, comparable to those in military conflicts.

As a result of disasters about 2 million people die annually in the world, more than 200 million suffer trauma of diverse severity, consequently about 10 million people remain disabled. 75-85 % of fatalities occur within first 20 minutes. According to some studies, the number of deaths would be reduced by 30 % if victims are provided medical care timely in an hour after catastrophe [1].

These figures demonstrate the importance of providing in proper time medical assistance on the disaster site.

In the case of emergency situation or disaster, resulting in a significant number of victims in a short period of time many people simultaneously require urgent medical assistance and evacuation from the impact zone. Inevitably, for a period of time there is a strain that the necessity in medical assistance exceeds currently available medical capabilities and resources. Obviously, in such circumstances medical aiding to the full extent for all affected people is practically impossible, that highlights the importance of emergency diagnostics, triage and setting a schedule for evacuation.

2 Triage in the Disaster Scenarios

Emergency situation is characterized by the complexity of decisions to be made.

Medical triage in case of mass casualty accidents or disasters [2, 3] is a complex process of identification and differentiation of victims in homogeneous groups according to the severity and nature of injuries and the degree of emergency medical assistance. It determines sequence, mode and evacuation destination depending on available medical capabilities and resources, as well as specific circumstances imposed by the impact.

The basic aim of medical triage is ensuring the provision of medical assistance in optimum time and in maximum possible volume to the largest number (ideally – to all) of victims of the disaster. In extreme cases diagnostics requires from physician a strategy that reduces to sorting victims into several groups in order to ensure targeted aid, taking into account the severity of the case.

Priority 1 – *Absolute emergency.* Victims with serious and very serious injuries, illnesses, intoxication or contamination, compromising vital functions, which require immediate stabilization measures, as well as priority evacuation in assisted medical transport conditions.

Priority 2 – *Relative emergency.* Victims with serious or moderate injuries, illnesses, intoxication or contamination, with retained vital functions, but with the risk of developing life-threatening complications immediately ahead. They require urgent medical assistance, but not the immediate one.

Priority 3 – *Low urgency.* Victims with minor injuries, illnesses, intoxication or contamination, no life-threatening, which can be treated later, usually in outpatient conditions. They can be evacuated in non-specialized transport or independently.

In case of mass casualty situations examination should be made on the site, in reduced time in order to determine the right strategy for saving. This specific character requires an approach, different from the traditional clinical examination. Structure of trauma and injuries varies essentially depending on the disaster's nature.

Ultrasound diagnostics at the disaster site is aimed at determining the level of urgency to save lives and to prevent any complications for people at risk.

Portable ultrasound has clear advantages over other imaging equipment (particularly, computed tomography) to be applied in remote places. Since ultrasound is painless and safe technique that captures images in real-time (showing the structure and movement of the body's internal organs and blood vessels) and portable light-weight ultrasound scanners can be used easily at the accident site, this method has been accepted as an important initial screening tool in disaster medicine.

For instance, to confirm liver injury ultrasound-competent physician has to answer the following four questions:

1. Is liver contour discontinued?

2. Are modifications in the liver structure (mostly circumscribed) observed?

3. Are collection(s) in the liver surrounding areas detected?

4. Does the patient have severe pains with acute onset (post-traumatic)?

If the answer is YES for all questions, then the obtained conclusion is "Post-traumatic lesions of liver with perihepatic hematoma", which corresponds to "Priority 1 – Absolute emergency".

FAST (Focused Assessment with Sonography for Trauma) examination, a well-known rapid bedside ultrasound examination used in emergency medicine, was grown widespread in the early 1990s.

"The FAST examination is based on the assumption that the majority of clinically significant abdominal injuries result in hemoperitoneum. The standard FAST protocol is directed to detection of fluid in the pericardial and peritoneal spaces. With regard to the CAVEAT protocol is limited to intraperitoneal hemorrhage" [4]. CAVEAT is the concept of a comprehensive sonographic examination in the evaluation of chest, abdomen, vena cava, and extremities in acute triage.

As discovery of free fluid in the abdomen can lead to appropriate and timely diagnostics, FAST can be used to guide clinical decision-making, e.g. as a quick method for triaging patients.

Rozycki et al. [5] have found that ultrasound was the most sensitive and specific in patients with penetrating chest wounds or in hypotensive blunt abdominal trauma patients (sensitivity and specificity nearly 100 %).

In [6] ultrasound was performed by relief teams after the 1988 Armenian earthquake as a primary screening procedure in 400 of 750 injured multiple casualty incident (MCI) patients admitted to a large hospital within 72 h of the event. The average time spent on evaluation of a single patient was approximately 4 min. Traumatic injuries of the abdomen were detected in 12.8 % of the patients.

In [7] sonography was used after an MCI in Guatemala in which the dead far outnumbered the injured. In that setting, ultrasound was useful for excluding internal trauma.

Ultrasound has been utilized in military deployments in Kosovo, Afghanistan and Iraq [8]. The British Air Assault Surgical Groups deployed to Kuwait during 2003 included the use of a hand-held ultrasound scanner by the forward medical units FAST examinations performed by trained emergency physicians using portable equipment at a large military combat hospital in Iraq. It had very high sensitivity as confirmed by subsequent operative reports and computed tomography imaging. In that particular experience ultrasound was performed in patients who sustained blunt, blast and penetrating trauma [8].

Statistics shows that in cases of natural disasters, catastrophes and accidents about 70 % of affected persons need specific healthcare approach limited in time.

So, it is important to offer recommendations on the evacuation priority and creation groups for evacuation from the disaster focus, according to destination (specialized centers or general profile medical institutions) based both on triage results and limited possibilities for transportation in case of a large number of victims.

The main limitation of the FAST examination is that the operator must be knowledgeable in its clinical use and be aware that *it does not*

exclude all injuries [9].

The limitations of FAST (as a task-focused approach) are caused by the narrow view in emergency situation formalization. In fact, the plausible reason of patient critical state can lie outside the emergency. For instance, the potential false-positive diagnosis of free traumatic fluid in the peritoneum may be due to fluid present in patients for physiologic reasons, including ovarian cyst rupture, as well as pathologic reasons, such as patients with ascites or inflammatory processes in the abdomen or pelvis [9].

In this paper we describe the algorithm that would allow physician rescue team in reduced time to appreciate diagnosis, severity, the life-threatening in order to make appropriate decisions about how to help, treat and evacuate from the disaster site.

In distinction from FAST, in our approach we consider emergency situation as a particular case of ultrasound examination domain formalization and take into account the injury severity.

3 Approach Based on SonaRes Knowledge Base

Abdomen region is the important one, as the most difficult cases to diagnose with extremely dangerous consequences are lesions of the abdominal cavity organs (liver, spleen, kidneys, large vessels of the abdomen, gallbladder and pancreas).

The paper authors over several years elaborated SonaRes technological platform, designed for development of medical informatics applications to support diagnostic process based on ultrasound examination method [10-11].

SonaRes technological platform consists of two main parts – SonaRes methodology and SonaRes technology. SonaRes methodology consists of knowledge acquisition strategy, knowledge representation and storage form, inference mechanism. SonaRes technology offers knowledge base editor and tools that allows creation of information systems of different destination.

The main part of SonaRes technological platform represents SonaRes knowledge base, which includes the following formalized expert knowl-

edge:

- 335 facts and 54 decision rules for gallbladder,
- 231 facts and 52 decision rules for pancreas,
- 167 facts and 31 decision rules for liver,
- 257 facts and 15 decision rules for bile ducts.

Based on facts, the decisions rules describe organs pathologies and anomalies.

Our approach presumes to re-engineer SonaRes knowledge base for on-site triage task in mass casualty situations, performing the following steps:

1. to add to the knowledge base information (facts and decision rules) that describes blood vessels;
2. to identify conclusions (decision rules) that describe fluid presence, obtaining *knowledge base – critical level 1*;
3. to localize the obtained conclusions into 4 areas of the FAST examination;
4. to identify information (facts and decision rules) that allows to make severity assessment (fluid volume and patient state severity), obtaining *knowledge base – emergency level 2*;
5. to identify conclusions (pathologies and anomalies) that describe presence of free fluid in the abdominal cavity, which is not the consequence of an abdominal injury, obtaining *knowledge base – non-injury level 3*;
6. to validate completeness of all 3 levels of the knowledge base for emergency situation;
7. to reorder the set of facts according to their information value in order to minimize the number of inference steps;

8. to classify conclusions from levels 1-2 in priority groups (absolute emergency, relative emergency, low urgency).

In this way we re-engineer SonaRes knowledge base structure from multimodule organ oriented to multilayered triage task oriented and overcome limitations of the FAST examination, taking into account physiologic and pathologic reasons.

4 Conclusion and Future Work

The current consensus supports ultrasound screening of mass casualties for evaluating trauma patients. In particular, FAST examination is used to identify presence of intraperitoneal or pericardial free fluid, presumed to be consequences of bleeding. It is important to note, however, that the FAST examination is a screening test, and false-negative conclusions do occur.

We propose a methodology of re-engineering of SonaRes knowledge base for on-site triage task in mass casualty situations.

As a result we obtain a multilayered knowledge base designed for emergency (mass casualty) situations, when injuries need immediate surgical intervention:

- *critical level 1* – corresponds to fluid presence
- *emergency level 2* – corresponds to severity assessment
- *non-injury level 3* – corresponds to presence of free fluid due to physiologic and pathologic reasons.

Our approach allows to differentiate process of on-site triage depending on time available for decision-making.

In cases when there is a need and the conditions allow (for instance, during transportation) to repeat examination, our approach, in distinction from FAST, provides more competent assistance, evaluating state severity assessment.

The following work could be done in the future:

- a scoring system to be added in priority groups (absolute emergency, relative emergency, low urgency), as it is usual for physicians;

- based on the re-engineered knowledge base, an algorithm to be created and validated on virtual scenarios.

In addition, there is a well suited provision of emergency physicians and rescue teams with a decision support system to assist emergency examination, helping in establishing the correct diagnostics in opportune terms. For example, it is possible to use SonaRes technological platform [10] that already exists and was tested in creation of a system that uses portable scanners and is aimed for diagnostics under field conditions, accessible through easy of use under mass casualty conditions, lack of time and qualified medical personnel.

References

[1] *Emergency medicine. Selected clinical lectures*, 3rd ed., vol. 1, V. V. Nikonov, A. E. Feskov, Eds. Donetsk: Publisher Zaslavsky A. Yu., 2008. (in Russian)

[2] J. L. Jenkins, M. L. McCarthy, L. M. Sauer, G. B. Green, S. Stuart, T. L. Thomas, E. B. Hsu, "Mass-casualty triage: Time for an evidence-based approach", *Prehospital Disaster Medicine*, vol. 23, no. 1, pp. 3–8, 2008.

[3] *Disaster Medicine*, 2nd ed., David E. Hogan DO, Jonathan L. Burstein MD, Eds. Lippincott Williams and Wilkins, a Wolters Kluwer business, 2016, 512 p.

[4] S. P. Stawicki, J. M. Howard, J. P. Pryor, D. P. Bahner, M. L. Whitmill and A. J. Dean, "Portable ultrasonography in mass casualty incidents: The CAVEAT examination," *World J. Orthopedics*, vol. 1, no. 1, pp. 10–19, 2010.

[5] G. S. Rozycki, R. B. Ballard, D. V. Feliciano, J. A. Schmidt and S. D. Pennington, "Surgeon-performed ultrasound for the assessment of truncal injuries: lessons learned from 1540 patients," *Ann Surg*, vol. 228, no. 4, pp. 557–567, 1998.

[6] A. E. Sarkisian, R. A. Khondkarian, N. M. Amirbekian, N. B. Bagdasarian, R. L. Khojayan and Y. T. Oganesian, "Sonographic screening of mass casualties for abdominal and renal injuries following the 1988 Armenian earthquake," *J Trauma*, vol. 31, no. 2, pp. 247–250, 1991.

[7] A. J. Dean, B. S. Ku and E. M. Zeserson, "The utility of handheld ultrasound in an austere medical setting in Guatemala after a natural disaster," *Am J Disaster Med*, vol. 2, no. 5, pp. 249–256, 2007.

[8] T. E. Kolkebeck and S. Mehta, "The focused assessment of sonography for trauma (FAST) exam in a forward-deployed combat emergency department: a prospective observational study," *Ann Emerg Med*, vol. 48, no. 4S, pp. 87–289, 2006.

[9] *AIUM Practice Parameter for the Performance of the Focused Assessment With Sonography for Trauma (FAST) Examination*, American Institute of Ultrasound in Medicine, 2014. [Online]. Available: www.aium.org/resources/guidelines/fast.pdf

[10] L. Burtseva, S. Cojocaru, C. Gaindric, O. Popcova and Iu. Secrieru, "Ultrasound diagnostics system SonaRes: structure and investigation process," in *Second International Conference "Modelling and Development of Intelligent Systems"*, Sibiu, Romania, September 29 – October 02, 2011, pp. 28–35.

[11] S. Cojocaru, C. Gaindric, O. Popcova and Iu. Secrieru, "SonaRes Platform for Development of Medical Informatics Applications," in *Proceedings of the 3rd International Conference on Nanotechnologies and Biomedical Engineering ICNBME-2015*, vol. 55. Chisinau, Moldova, September 23-26, 2015, pp. 450–453.

Svetlana Cojocaru, Constantin Gaindric,
Iulian Secrieru, Sergiu Puiu, Olga Popcova

Received July 25, 2016

Svetlana Cojocaru
Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova
5 Academiei street, Chisinau, MD 2028
E–mail: svetlana.cojocaru@math.md

Constantin Gaindric
Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova
5 Academiei street, Chisinau, MD 2028
E–mail: gaindric@math.md

Iulian Secrieru
Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova
5 Academiei street, Chisinau, MD 2028
E–mail: secrieru@math.md

Sergiu Puiu
Women's Health Center "Dalila"
E–mail: puiusv@yahoo.com

Olga Popcova
Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova
5 Academiei street, Chisinau, MD 2028
E–mail: oleapopcova@yahoo.com

Computer Science Journal of Moldova, vol.**24**, no.2(71), 2016

The Chromatic Spectrum of a Ramsey Mixed Hypergraph

David Slutzky, Vitaly Voloshin

Abstract

We extend known structural theorems, primarily a result of Axenovich and Iverson, for the strict edge colorings of the complete graph K_n which avoid monochromatic and rainbow triangles to discover recursive relationships between the chromatic spectra of the bihypergraphs modeling this coloring problem. In so doing, we begin a systematic study of coloring properties of mixed hypergraphs derived from coloring the edges of a complete graph K_n in such a way that there are no rainbow copies of K_r and no monochromatic copies of K_m, where $n \geqslant r \geqslant 3$, $n \geqslant m \geqslant 3$. We present the chromatic spectra of the bihypergraph models of K_n for $4 \leqslant n \leqslant 12$ and $r = m = 3$. This study fits in the larger context of investigating mixed hypergraph structures that realize given spectral values, as well as investigations of the sufficiency of the spectral coefficients in obtaining recursive relationships without the need to subdivide them further into terms that count finer distinctions in the feasible partitions of the hypergraph. The bihypergraphs arising in this simplest case where $r = m = 3$ have spectra that are gap free and which do allow a recursive relationship, albeit a complicated one. The continuation of this project in future work will examine if both of these facts remain true for derived Ramsey Mixed Hypergraphs corresponding to larger r and m.

Keywords: Ramsey number, antiramsey number, mixed hypergraph coloring, feasible partition, chromatic spectrum.

Mathematics Subject Classification: 05C15, 05C65

1 Definitions

A *hypergraph* $\mathcal{H} = (X, \mathcal{E})$ is a collection of vertices $X = \{x_i | i \in I\}$ and a collection of hyperedges $\mathcal{E} = \{e_j \subseteq X | j \in J\}$. A *coloring* of $\mathcal{H}$ is a mapping $f : X \to [k]$, where $[k] = \{1, ..., k\}$. The inverse image $f^{-1}(i)$ is a *color set* defined by the coloring f and the collection of the nonempty color sets defined by f gives a partition of X. When f is a surjection, the coloring is said to be *strict*. Below, all colorings are assumed to be strict unless stated otherwise. In *mixed hypergraph coloring* ([14]–[16]) there are subsets $\mathcal{C}$ and $\mathcal{D}$ of the hyperedge set $\mathcal{E}$ which place conditions on colorings. A coloring f is *proper* if it assigns the same color to at least two vertices in each hyperedge in $\mathcal{C}$ and different colors to at least two vertices in each hyperedge in $\mathcal{D}$. Hence, a hyperedge in $\mathcal{C}$ cannot be rainbow (all vertices of distinct colors), and a hyperedge in $\mathcal{D}$ cannot be monochrome (all vertices of the same color). For convenience, we refer to the hyperedges in $\mathcal{C}$ and $\mathcal{D}$ as C-edges and D-edges. The partitions of X corresponding to proper colorings are the *feasible partitions* of $\mathcal{H}$. A mixed hypergraph $\mathcal{H} = (X, \mathcal{C}, \mathcal{D})$ is a *bihypergraph* when $\mathcal{E} = \mathcal{C} = \mathcal{D}$, thereby requiring both coloring conditions on every hyperedge of $\mathcal{H}$.

Mixed hypergraph coloring has many diverse applications. The monograph [16] gives an overview of many of these applications, such as list colorings without lists and problems in resource allocation, data base management, and molecular biology. As it is shown in [10], mixed hypergraphs can be used to efficiently model many graph coloring problems including homomorphisms of simple graphs and multigraphs; circular colorings; (H, C, K)-colorings; locally surjective, locally bijective, and locally injective homomorphisms; $L(p, q)$-labelings; the channel assignment problem; and T-colorings and generalized T-colorings. There are also applications of mixed hypergraph coloring to issues regarding cybersecurity. One such reference is a recent PhD thesis [12] giving algorithms for scalable fault tolerance.

It is well-known, see [16], that the chromatic polynomial $P = P(\mathcal{H}) = P(\mathcal{H}, \lambda)$, which is the function that counts the number of, not necessarily strict, proper λ-colorings of $\mathcal{H}$, can be expressed in the

form

$$P = \sum_{i=1}^{n} R_i \lambda^{\underline{i}}, \tag{1}$$

where $|X| = n$ and $\lambda^{\underline{i}} = \lambda(\lambda - 1) \cdots (\lambda - i + 1)$ is the falling factorial. Note that the falling factorial counts the number of colorings of the complete graph on i vertices. The coefficients R_i in (1) count the number of feasible partitions of $\mathcal{H}$ using i nonempty subsets. Without coloring conditions, given by defining the sets $\mathcal{C}$ and $\mathcal{D}$ in $\mathcal{E}$, R_i is the Stirling number of the second kind $S(n, i)$ and their sum is the nth Bell number B_n counting the number of partitions of a set of order n. See, for example, [7]. With coloring conditions, we have the trivial bound $R_i \leqslant S(n, i)$. The collection of these coefficients $\{R_1, ..., R_n\}$ is called the *chromatic spectrum* of $\mathcal{H}$. The smallest value of i for which R_i is nonzero is the (lower)-chromatic number $\chi(\mathcal{H})$ and the largest value of i for which R_i is nonzero is the upper-chromatic number $\overline{\chi}(\mathcal{H})$. If all R_i are nonzero for $\chi(\mathcal{H}) \leqslant i \leqslant \overline{\chi}(\mathcal{H})$, the spectrum is said to be *gap free*. The set of indices for which the spectral coefficients are nonzero is the *feasible set* of $\mathcal{H}$. Not only are gaps possible, but any finite set of positive integers is the feasible set of some mixed hypergraph if and only if the set omits 1, or includes 1 and is gap free. See [8].

The splitting-contraction algorithm [14]–[16] finds the chromatic polynomial of any mixed hypergraph, but it has a high level of computational complexity that makes it impractical to use in large hypergraphs. In [13], we use an extension to the splitting-contraction algorithm in the special case of complete uniform interval mixed hypergraphs to find recursive relationships between their chromatic polynomials, where the recursion is on the order of their vertex sets. Recursive relationships for the chromatic polynomials naturally include recursive relationships for the chromatic spectrum values. Our main results here regard a collection of bihypergraphs which model a particular question in Ramsey Theory. We find recursive relationships for the individual chromatic spectral values of these bihypergraphs directly, and do not consider further the full chromatic polynomials. By focusing on the chromatic

spectra, we are focused on the growth patterns of the collections of feasible partitions. We give two different ideas of equivalence of colorings to distinguish between the objects counted by the chromatic spectral values and coloring patterns that are the same in a weaker sense.

Two colorings f and g are *isomorphic* if there is a permutation σ of the vertex set X so that $f = g \circ \sigma$. Two k-colorings are *equivalent* if there is a permuation θ of the colors $[k]$ so that $f = \theta \circ g$. Equivalent colorings give the same feasible partition, whereas isomorphic colorings only give the same coloring pattern. Since the chromatic spectral values count the number of nonequivalent colorings, or isomorphic colorings with multiplicity, we say nonequivalent colorings are distinct.

2 Background of a Ramsey Problem

Let $f : \mathcal{E}(K_n) \to [k]$ be an edge coloring of the complete graph on n vertices. Let G_m and G_r be two graphs. The coloring f is said to be (G_m, G_r)-good if there is no monochrome subgraph isomorphic to G_m and no rainbow subgraph isomorphic to G_r. Define the derived *Ramsey Mixed Hypergraph* $\mathcal{H}_R = (\mathcal{E}(K_n), \mathcal{C}(G_r), \mathcal{D}(G_m))$ to be the mixed hypergraph with vertex set corresponding to the edge set of K_n, C-edges corresponding to the copies of G_r in K_n, and D-edges corresponding to the copies of G_m in K_n. In a more general setting this concept was first introduced by Voloshin in 2002 in [16, p.157] under the name of "derived mixed hypergraph of a hypergraph $\mathcal{H} = (X, \mathcal{E})$". In this language, for example, the classic graph Ramsey number $R(p, p)$ (the smallest integer n such that any 2-coloring of $\mathcal{E}(K_n)$ contains a monochromatic copy of K_p in color 1 or a monochromatic copy of K_p in color 2) is the smallest integer n such that the lower chromatic number $\chi(\mathcal{H}_R) = \chi(\mathcal{E}(K_n), \emptyset, \mathcal{D}(K_p)) > 2$.

Clearly, a (G_m, G_r)-good edge coloring f of K_n is a proper coloring of the mixed hypergraph $\mathcal{H}_R$. Axenovich and Iverson [1] define $maxR(n; G_m, G_r)$ and $minR(n; G_m, G_r)$ to be the maximum and minimum number of colors, respectively, in a (G_m, G_r)-good edge coloring of K_n. These numbers are the upper and lower chromatic numbers of $\mathcal{H}_R$. Further, let $F(k; G_m, G_r)$ be the largest value of n for which there

is a (G_m, G_r)-good edge k-coloring of K_n. These kinds of numbers have been studied by many authors in Ramsey and anti-Ramsey Theory, see for example [3]–[5], [9].

For the remainder we work with the particular case when $G_m \cong G_r \cong K_3$. As such, it should be understood that good colorings in the following are (K_3, K_3)-good edge colorings of a complete graph. By an easy induction argument, it can be seen that $maxR(n; K_3, K_3) = n-1$. In [6] the authors show that every good $(n-1)$-coloring can be obtained as a kind of product of two good colored cliques. In [3], Chung and Graham prove that

$$F(k; K_3, K_3) = \begin{cases} 5^{k/2} & \text{if } k \text{ is even} \\ 2 \cdot 5^{(k-1)/2} & \text{if } k \text{ is odd} \end{cases} \tag{2}$$

by examining the *monochrome neighborhoods* of a fixed vertex x, defined by $N_i(x) = \{y | f(xy) = i\}$. They also remark that a similar analysis shows that the colorings in the extremal cases can be described using a recursive process with two kinds of products on colored cliques with two and five factors. Axenovich and Iverson in [1] give a more detailed account of these extremal colorings by using product structures on sets of colorings that equate to the products described by Chung and Graham. Axenovich and Iverson also give a proof examining monochrome neighborhoods, and use a structural lemma begun in an earlier paper by Axenovich and Jamison [2]. In the next section we state their lemma and show that it, in fact, gives a description of all good edge colorings of K_n using the same kind of product structures, but we need a product with four factors in addition to the products with two and five factors. Note that we favor the perspective of products of colored subgraphs, as described by Chung and Graham, as opposed to the products of sets of colorings in Axenovich and Iverson. These products are built by designating color patterns on the join edges of an underlying join of the factors. Though we refer to these as product structures, they are actually multivalued products that are more clearly described as blow ups of the coloring patterns used on the join edges. That perspective was taken by Axenovich and Iverson in their products of sets of color-

ings. We explicitly define these products after listing the isomorphism classes of the good colorings for K_2 through K_5.

The structural theorem on the good edge colorings gives recursive relationships between the chromatic spectral values of the corresponding bihypergraphs. The recursion is on n, the size of the underlying complete graphs which generate this family of derived 3-regular bihypergraphs. Many authors have examined complexity issues and robustness of mixed hypergraphs, and it is known that even 3-regular bihypergraphs provide diverse models with a lot of complexity. Even the computation of the chromatic spectrum of 3-regular bihypergraphs is a hard problem. For example, in [13], the bihypergraph case for uniform complete interval hypergraphs, which have a relatively simple structure, is shown to be much harder than the cases where all the hyperedges are $\mathcal{C}$-edges or all $\mathcal{D}$-edges. In the latter cases simple recursive relationships are found for all sizes of the uniform edges, but for bihypergraphs only recursive relationships are shown for bi-edges of size 3 or size 4. Our main result is the recursive relationships that follow from the product structure. However, we can also use the product structure to find an explicit formula for the leading coefficient, and indicate some weak bounds on the growth of the spectral coefficient of degree $n-2$. After commenting on the techniques used to obtain these results, we conclude with the chromatic spectra corresponding to K_2 through K_{12}, a description of the Java program used to compute these data using the recursive relationship, and a description of a brute force sorting algorithm that was also used to count feasible partitions up through the K_{11} case.

3 Structure of Good Colorings

In Figure 1 we list the isomorphic colorings, or coloring patterns, for all of the good colorings of K_2 through K_5. With each figure we list the number of distinct colorings in that isomorphism class.

The coloring patterns in Figure 1 can be obtained by hand, though it can be tedious work to get the patterns for K_5. They also result from

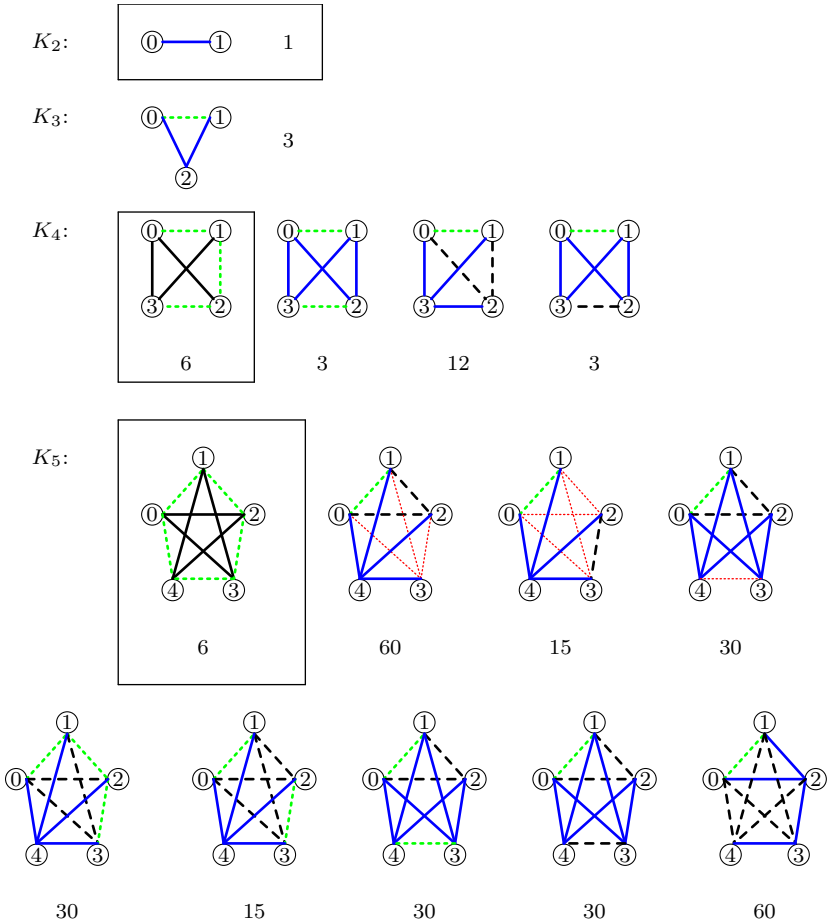

Figure 1. The isomorphism classes of (K_3, K_3)-good coloring patterns for K_2 through K_5 and the number of distinct colorings in each class. The designated patterns are the prime patterns used to define products that generate all good colorings of any K_n.

the forthcoming structural theorem. The three patterns in Figure 1 in the three boxes are the *prime* patterns used to define the necessary product structures.

Let the symbol $[r_0, r_1]$ represent any of the colored complete graphs obtained by replacing the vertices in the prime coloring of K_2 with colored complete graphs K_{r_0} and K_{r_1} where the factors are edge colored with colors different from that used in the original K_2. The resulting complete graph $K_{r_0+r_1}$ is the join of the two underlying factors with each of the join edges colored by the color in the original K_2. This is the operation used in [6] and is equivalent to operations appearing in [3] and [1].

Let the symbol $[r_0, r_1, r_2, r_3]$ represent any of the colored complete graphs obtained by replacing the vertices in any of the six prime colorings of K_4 with colored complete graphs $K_{r_0}, K_{r_1}, K_{r_2}$, and K_{r_3}, where the factors are edge colored with colors different from the two used in the original K_4.

Let the symbol $[r_0, r_1, r_2, r_3, r_4]$ represent any of the colored complete graphs obtained by replacing the vertices in any of the six prime colorings of K_5 with colored complete graphs $K_{r_0}, K_{r_1}, K_{r_2}, K_{r_3}$, and K_{r_4}, where the factors are edge colored with colors different from the two used in the original K_5. This is the same operation as the one defined in [3], but they used just one of the six 2-colorings of K_5 and considered permutations of the factors to obtain the other results of our multivalued product.

For convienence, when the orders of the factors are known, we always order the factors so that $r_0 \geqslant r_1 \geqslant \ldots \geqslant r_4$.

The products of sets of colorings defined in [1] give all of these products and more, but they only use the ones corresponding to these two and five factor products to describe the extremal coloring patterns corresponding to (2).

The three distinct good colorings of K_3 are obtained from the product $[2, 1]$. The nonprime good colorings of K_4 are obtained from $[2, 2]$ and $[3, 1]$. Note that the operation $[2, 2]$ gives 2-colorings when the same color is used on the factors and 3-colorings when different colors are used on the factors. The first two nonprime good colorings of K_5

shown in Figure 1 are 4-colorings obtained from using the two nonisomporphic 3-colorings of K_4 in $[4,1]$. The last good 4-coloring pattern is obtained from $[3,2]$. The five good 3-coloring patterns of K_5 are given by $[4,1],[4,1],[3,2],[3,2]$, and $[2,1,1,1]$, respectively. The two nonisomorphic good 4-coloring patterns given by $[4,1]$ come from the two nonisomorphic good 2-colorings of K_4. However, the two nonisomorphic good 4-coloring patterns given by $[3,2]$ come from the two choices of colors on K_2. See Section 4 to see feasible partitions of the labeled edge set of K_5 corresponding to these coloring patterns.

We now show that these three operations are enough to generate all of the good colorings for any K_n.

Following Axenovich and Iverson, we define a *monochromatic pair* and a *mixed pair* of subsets of vertices of an edge colored complete graph. Let f be a good coloring of K_n and let A and B be subsets of $V(K_n)$. Put $c(A,B)$ equal to the set of colors f assigned to edges ab for any $a \in A$ and $b \in B$. A pair (A,B) is monochromatic if $c(A,B) = \{i\}$. The pair is mixed if there are partitions $A = A' \cup A''$ and $B = B' \cup B''$ with $c(A',B') = c(B',B'') = c(B'',A'') = i$ and $c(A',B'') = c(A',A'') = c(A'',B') = \{j\}$. The sets A' and B' may be empty, but not both. See Figure 2.

Lemma 1. *[Axenovich/Iverson] Let $v \in V(K_n)$ and $V_i = N_i(v)$. Order the colors so that $V_i \neq \emptyset$ for $i = 1,\ldots,m$. Then there is a reordering such that:*

a) $c(V_i,V_j) = \{i,j\}$ and the pair (V_i,V_j) is either monochromatic or mixed,

b) If (V_i,V_j) is monochromatic, then $c(V_i,V_j) = \{i\}$,

c) If (V_i,V_j) is mixed, then $j = i+1$,

d) If there is a mixed pair (V_i,V_j), then neither (V_{i-1},V_i) or (V_{i+1},V_{i+2}) is mixed.

Figure 2. A mixed pair A and B.

An edge coloring is *lexical* if there is an ordering of the vertex set so that each edge v_iv_j is assigned color i whenever $i < j$. As Axenovich and Iverson comment, their lemma classifies good colorings as blow ups of lexical colorings, with the possible exceptions of mixed pairs of consecutive sets. A set can be in only one mixed pair. See Figure 3.

Theorem 1. *Any good colored K_n is obtained uniquely as one of the products from $[r_0, r_1]$, $[r_0, r_1, r_2, r_3]$, or $[r_0, r_1, r_2, r_3, r_4]$, with $r_0 \geqslant r_1 \geqslant \ldots \geqslant r_4$ and with good-colorings on each factor.*

Proof Choose a vertex $v \in V(K_n)$ and order the monochrome neighborhoods of v as in the conclusion of Lemma 1.

Suppose $m = 1$. Then $V(K_n) - v = N_1(v)$ and the colored K_n is realized by $[n-1, 1]$ with the join edges colored by color 1.

Suppose $m \geqslant 2$. There are two cases to consider. See Figure 3.

If (V_1, V_2) is a monochromatic pair, then $c(V_1, V_j) = \{1\}$ for all $j > 1$ and the colored K_n is realized by a product with two factors given by the induced subgraphs $K_n[v \cup \bigcup_{i=2}^{m} V_i]$ and $K_n[V_1]$.

If (V_1, V_2) is a mixed pair, then $c(V_1, V_j) = \{1\}$ for all $j > 2$, $c(V_2, V_j) = \{2\}$ for all $j > 2$, and the colored K_n is realized by a product with four or five factors, depending on whether one of the subsets

of the partitions of V_1 or V_2 is empty. The factors are the subgraphs induced by V_1', V_1'', V_2', V_2'', and $v \cup \bigcup_{i=3}^{m} V_i$.

The uniqueness follows immediately, since the colors on the join edges of the products are incident to every vertex in each of the three product structures.

□

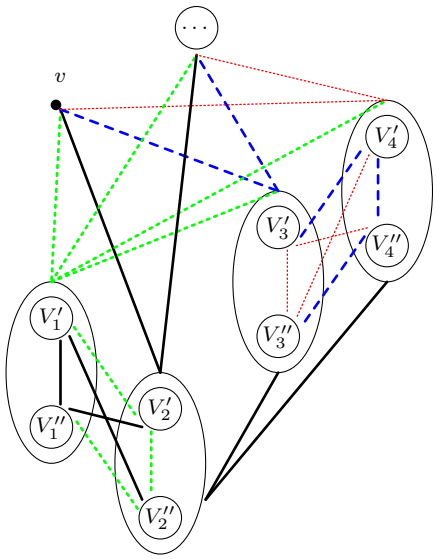

Figure 3. The product structures of any good coloring for Theorem 1. When V_1' and V_2' are both empty, the product has two factors. When V_1' is empty, the product has four factors. Otherwise the product has five factors.

There are no restrictions on the size of the factors in Theorem 1. When each factor is K_1, we get the three prime coloring patterns of Figure 1. Considering the possible arithmetic partitions of 5 of sizes

2, 4, and 5, the patterns shown in Figure 1 form a complete list of the good coloring patterns for K_5.
If we color each factor using one fewer color than its order, we have that each product can have up to $(r_0 - 1) + (r_1 - 1) + (1) = n - 1$, $(r_0-1)+(r_1-1)+(r_2-1)+(r_3-1)+(2) = n-2$, or $(r_0-1)+(r_1-1)+(r_2-1)+(r_3-1)+(r_4-1)+(2) = n-3$ colors. The full structural theorem then immediately gives the result of [6] that the good colorings using the maximum number of colors $n-1$ are only obtained from products with two factors with the maximum number of colors used on each factor and these sets of colors are distinct. The minimum number of colors is the maximum of the lower chromatic numbers corresponding to the factors plus 1 if there are two factors, or plus 2 if there are four or five factors. These minimums clearly cannot decrease as the order increases. Further, it is inductively clear, using products with two factors, that the spectra of these derived bihypergraphs are gap free. We are interested to see if any of the derived Ramsey Mixed Hypergraphs, for some combination of K_r and K_m, realize gaps in their chromatic spectra. We plan to investigate that possibility in future research continuing the project begun in this paper.

Define R_i^n to be the coefficients of the chromatic spectrum of the derived Ramsey Bihypergraph of K_n with bi-edges corresponding to the triangles of K_n.

Theorem 2. *The leading coefficient R_{n-1}^n is $(2n-3)!!$*

Theorem 2 follows immediately from the following lemma.

Lemma 2. *Each distinct good $(n-2)$-coloring of K_{n-1} extends to $(2n-3)$ distinct good $(n-1)$-colorings of K_n.*

Proof It is easy to check that each of the 3 distinct good 2-colorings of K_3 extends to 5 distinct good 3-colorings of K_4. Now assume the

statement of the lemma is true for all orders up through $n-1$. Hence, each good $(r-1)$-coloring of K_r extends to $2(r+1)-3 = 2r-1$ good r-colorings of K_{r+1} for $r = 1, \ldots, (n-2)$. Pick any $v \in V(K_n)$ and any good $(n-2)$-coloring of K_{n-1} applied to $K_n - v$. From the structural theorem and comment following it, this coloring can be obtained as a product $[r, s]$ with $r + s = n - 1$ using $r - 1$ colors on the first factor and $s - 1$ distinct colors on the second factor. For convenience let's say the color on these join edges is green. To increase the number of colors used on K_n, at least one of the edges joining v to K_{n-1} must be a new color, say red. To avoid creating rainbow triangles, there are two possibilities: all of these edges can be red, or some of these edges joining v to one of the factors K_r or K_s are red while all of the edges joining v to the other factor are green. From the induction hypothesis we see there are $1 + (2r-1) + (2s-1) = 2n-3$ possible good colorings of K_n produced from the selected good $(n-2)$-coloring of K_{n-1}.

□

The reasoning in the proof of Lemma 2 extends to products with four and five factors, however we can only obtain a much weaker result than that of Theorem 2.

Theorem 3. *The coefficient R^n_{n-2} is greater than $(2n-5)R^{n-1}_{n-3}$.*

Proof Theorem 3 is equivalent to the statement: each distinct good $(n-3)$-coloring of K_{n-1} extends to at least $(2n-5)$ distinct good $(n-2)$-colorings of K_n.

We begin the induction with K_5. The prime 2-colorings of K_4 each extend to 5 good 3-colorings of K_5 when we join a single vertex to K_4. However, each of the nonprime 2-colorings of K_4 extends to 7 good 3-colorings of K_5. Without subdividing this collection of good colorings, we have the weak relationship that $R^5_3 > 5R^4_2$. Note we are also ignoring any 3-colorings of K_5 that are extensions of good 3-colorings of K_4. Now assume the statement of the claim is true for all orders up through $n-1$. Pick any $v \in V(K_n)$ and any good $(n-3)$-coloring of K_{n-1} applied to $K_n - v$. From the structural theorem and comment following it, this coloring can be obtained as a product with two factors or as a product with four factors.

Treating first the cases when the coloring on K_{n-1} is a product with four factors, we again note that each of the factors must be colored with the maximum number of colors r_i-1, and there must be a new color, say red, used on at least one of the join edges leading to v when we extend this coloring to a coloring of K_n. It could be the case that all of the edges joining v to K_{n-1} are red. Otherwise, let's begin by assuming the red edges lead to only one factor, say K_{r_0}. To avoid rainbow triangles, the edges joining v to another factor, say K_{r_1}, must all be colored using the color on the edges joining the two factors K_{r_0} and K_{r_1}. Since we can replace r_1 with r_2 or r_3 in the previous observation, we also see that these are in fact the only possible patterns; either all of the edges are red, or there are only red edges leading to one factor and the edges leading to the other factors are determined by the product structure. As in the proof of Lemma 2, this good $(n-3)$-coloring on K_{n-1} extends to $2r_0-1+2r_1-1+2r_2-1+2r_3-1+1 = 2n-5$ good $(n-2)$-colorings on K_n.

In the cases when the good coloring on K_{n-1} is a product with two factors, and noting that we are requiring a new color be used on at least one edge joining v to K_{n-1}, there are three subcases possible. Either one of the factors is colored with the maximum number of colors $r_i - 1$ while the other is colored with one fewer than the maximum number, $r_j - 2$, and the colors used on the factors are distinct; or both factors are colored with the maximum number of colors but share one color. In the first subcase we use the inductive hypothesis that each good $(r_i - 2)$-coloring of K_{r_i} extends to at least $(2r_i - 3)$ distinct good (r_i-1)-colorings of K_{r_i+1} to get that this good $(n-3)$-coloring of K_{n-1} extends to at least $1+2r_i-3+2r_j-1 = 2n-5$ good $(n-2)$-colorings of K_n. In the last subcase, we have that the good coloring extends to at least $1 + 2r_0 - 1 + 2r_1 - 1 = 2n - 3$ good colorings. Not only is $2n - 3$ greater than $2n - 5$, but particular colorings that use the shared color on all of the edges incident with one vertex of each of the two factors permit additional extensions with edges joining v to these vertices possibly colored with this shared color. This completes the proof.

□

Though any good coloring of K_n must be an extension of a good coloring of K_{n-1}, Theorem 3 shows the need to analyze the details of a particular coloring pattern on K_{n-1} to determine how many extensions it allows. That analysis is even more complex when we do not increase the number of colors used. Chung and Graham show, in obtaining (2), that if we do not increase the number of colors used, the number of good colorings increases for a while and then decreases to zero. In [13] we first subdivide each spectral coefficient into terms correpsonding to individual patterns and then piece the subdivisions together to find recursive relationships between the spectral coefficients themselves. In this case, with these derived bihypergraphs, the product structure with only three prime patterns immediately produces a recursive relationship between the spectral coefficients. The details of the contributions made by different coloring patterns are inherent in this relationship with the repetition of all of the coefficients throughout the formula. This formula permits relatively efficient computation of the spectral coefficients, but does not easily permit more insight, of the form given above, of the growth patterns of these values.

To state the general recursions we need more terms. Since the recursion is very large, we use a single sigma representing summations over multiple parameters.

Let P_2, P_4, and P_5 be the sets of arithmetic partitions of n of size $2, 4$, and 5 with terms arranged in nonincreasing order. Elements of these sets form the products $[r_0, r_1]$, $[r_0, r_1, r_2, r_3]$, and $[r_0, r_1, r_2, r_3, r_4]$. Let $D_2^{[r_0,r_1]}, D_4^{[r_0,r_1,r_2,r_3]}$, and $D_5^{[r_0,r_1,r_2,r_3,r_4]}$ be the number of ways of decomposing K_n as the join of the factors K_{r_i} in the respective case. The numbers D_i^* are functions of a particular arithmetic partition and are multinomial coefficients divided by factorials corresponding to numbers of repeated terms in the partition.
We use $C_k^n = n!/k!/(n-k)!$ for the binomial coefficient n choose k and $P_k^n = n!/(n-k)!$ for the number of permutations of k objects chosen from n objects.

Theorem 4. *The general coefficient R_c^n is given by*

$$D_2^{[n-1,1]}R_{c-1}^{n-1} + 6D_4^{[n-3,1,1,1]}R_{c-2}^{n-3} + 6D_5^{[n-4,1,1,1,1]}R_{c-2}^{n-4} +$$
$$\sum_{P_2,r_1>1} D_2^{[r_0,r_1]}R_p^{r_0}R_q^{r_1}C_{z_1}^p P_{z_1}^q +$$
$$\sum_{P_4,r_1>1,r_2=1} 6D_4^{[r_0,r_1,1,1]}R_p^{r_0}R_q^{r_1}C_{z_2}^p P_{z_2}^q +$$
$$\sum_{P_4,r_2>1,r_3=1} 6D_4^{[r_0,r_1,r_2,1]}R_p^{r_0}R_q^{r_1}R_r^{r_2}C_u^p P_u^q C_w^{t_1} P_w^r +$$
$$\sum_{P_4,r_3>1} 6D_4^{[r_0,r_1,r_2,r_3]}R_p^{r_0}R_q^{r_1}R_r^{r_2}R_s^{r_3}C_u^p P_u^q C_v^{t_1} P_v^r C_x^{t_2} P_x^s +$$
$$\sum_{P_5,r_1>1,r_2=1} 6D_5^{[r_0,r_1,1,1,1]}R_p^{r_0}R_q^{r_1}C_{z_2}^p P_{z_2}^q +$$
$$\sum_{P_5,r_2>1,r_3=1} 6D_5^{[r_0,r_1,r_2,1,1]}R_p^{r_0}R_q^{r_1}R_r^{r_2}C_u^p P_u^q C_w^{t_1} P_w^r +$$
$$\sum_{P_5,r_3>1,r_4=1} 6D_5^{[r_0,r_1,r_2,r_3,1]}R_p^{r_0}R_q^{r_1}R_r^{r_2}R_s^{r_3}C_u^p P_u^q C_v^{t_1} P_v^r C_x^{t_2} P_x^s +$$
$$\sum_{P_5,r_4>1} 6D_5^{[r_0,r_1,r_2,r_3,r_4]}R_p^{r_0}R_q^{r_1}R_r^{r_2}R_s^{r_3}R_t^{r_4}C_u^p P_u^q C_v^{t_1} P_v^r C_j^{t_2} P_j^s C_y^{t_3} P_y^t,$$

where $t_1 = p+q-u$, $t_2 = p+q-u+r-v$, $t_3 = p+q-u+r-v+s-j$, $z_1 = p+q-c+1$, $z_2 = p+q-c+2$, $w = t_1+r-c+2$, $x = t_2+s-c+2$, $y = t_3+t-c-2$, and each of the sums is over all possible choices of colors on the factors so that the total number of colors used is $c-2$ (except the first sum which has $c-1$ colors) and over each of the possible numbers of intersections of colors between the factors.

Proof We only explain the last summation, as the earlier terms follow from a similar counting argument, where we must separate cases based on how many factors of each product are K_1. The last sum gives the good c-colorings obtained from products with five factors which

are all bigger than K_1. The symbol D_5^* is the number of ways of decomposing K_n with these factors, and there are six distinct coloring patterns possible on the edges joining these factors for each decomposition and each of the possible coloring patterns on the individual factors. We must multiply by the numbers of distinct good colorings of the factors with p, q, r, s, and t colors used on each of the factors, where $p + q + r + s + t \geqslant c - 2$. The first two factors may share u colors. If so, we must multiply by the number of ways of choosing the common u colors from the p colors used on the first factor when we color the second factor. Moreover, we get different coloring patterns by permuting these common colors amongst themselves in the coloring of the second factor. The third factor may share v colors with the union of the first two factors. The fourth factor may share j colors with the union of the first three factors. The term y is the number of colors the fifth factor must share with the previous four factors, given the particular choices made on the colors of the previous factors.

□

In the context of finding the chromatic polynomial, or chromatic spectrum, the above recursion is interesting for the complexity of the recursion that can exist even for a 3-regular bihypergraph. Perhaps most interesting are the factors of 6 that represent a subset of the good 2-colorings and not the entire spectral value R_2^4. Are there other examples of ordered families of mixed hypergraphs which exhibit recursive relationships which require a finer decomposition than that which is given by the full chromatic spectrum? In the example given above, that finer level of decomposition only required subdividing one set of feasible partitions. Are there other families that require subdividing more of the spectral values in recursive relationships? In [13] we conjecture that this may be the case for uniform complete interval bihypergraphs with edge size bigger than 4.

4 The spectra for K_2 through K_{12}

Corollary 1. *The chromatic spectrum values for the derived Ramsey Bihypergraphs corresponding to K_2 through K_{12} are as follows:*

K_2: $[1]$
K_3: $[0, 3]$
K_4: $[0, 9, 15]$
K_5: $[0, 6, 165, 105]$
K_6: $[0, 0, 846, 2790, 945]$
K_7: $[0, 0, 3402, 42273, 49770, 10395]$
K_8: $[0, 0, 10836, 557928, 1604925, 961695, 135135]$
K_9: $[0, 0, 19278, 6972966, 44972172, 55829655, 20210715, 2027025]$
K_{10}: $[0, 0, 14742, 81569754, 1201982166, 2778115725, 1906370235, 460971000, 34459425]$
K_{11}: $[0, 0, 0, 875500164, 31404779406, 130507877808, 151891001955, 65874757410, 11365685100, 654729075]$
K_{12}: $[0, 0, 0, 8588423844, 808974051732, 6002106197472, 11277243566820, 7842349288620, 2338009242030, 301618981125, 13749310575]$

The data in Corollary 1 were computed using Java following the recursive relationships in Theorem 4. After first generating the sets of arithmetic partitions P_2, P_4, and P_5 and computing the number of decompositions D_2, D_4, and D_5 for each such partition, one follows the cases of the recursion to compute the number of distinct good colorings contributed by each arithmetic partition. This program can be extended to find the spectra corresponding to larger K_n using big integers to avoid errors created due to loss of memory in Java with long variables after K_{12}. Using longs, Java takes less than 2 seconds (on a Dell desktop with an Intel(R) Core(TM) i7 CPU 860 processor) to compute these spectral values.

We also computed most of the data in Corollary 1 using brute force sorting algorithms to count all good colorings via feasible partitions. Once labels and an ordering of the edges of K_n are chosen, restricted growth strings encode the feasible partitions, see [11] for instance. We

choose to label the edges lexically, which means we first label all the edges incident to vertex 0 and then the remaining edges incident to vertex 1 etc. To illustrate this ordering, the distinct good edge colorings of K_5 used as the representatives of the isomorphism classes shown in Figure 1 correspond to the following feasible partitions:

$$\{\{0,3,4,7,9\},\{1,2,5,6,8\}\}$$

$$\{\{0,4,7\},\{1,2,5\},\{3,6,8,9\}\},\ \{\{0,7\},\{1,2,4,5\},\{3,6,8,9\}\},$$
$$\{\{0,9\},\{1,4\},\{2,3,5,6,7,8\}\},\ \{\{0\},\{1,4,9\},\{2,3,5,6,7,8\}\},$$
$$\{\{0\},\{1,4,7,9\},\{2,3,5,6,8\}\}$$

$$\{\{0\},\{1,4\},\{2,5,7\},\{3,6,8,9\}\},\ \{\{0\},\{1,2,4,5\},\{3,6,8,9\},$$
$$\{7\}\},\{\{0\},\{1,4\},\{2,3,5,6,7,8\},\{9\}\}$$

By labeling edges lexically, one can easily determine the labels on all of the triangles. Generate the first $n-1$ entries of the growth string. Upon the generation of the nth entry we encounter a condition on the strings due to a triangle. We then continue with the process only if this condition, that exactly two colors are used, is met. As we generate each additional term of the string we encounter more triangles and only continue the generation of the string if the corresponding conditions are met. The number of triangles encountered at each stage of the string generation goes up by one once the first index on the edge increases. Strings are counted if they make it to the end of the generation process based on their highest entry, and these counts give the chromatic spectral values. Using the same machine as above, it took 3.5 hours to run the sorting algorithm for K_{10}, and it took 75 days to run the sorting algorithm for K_{11}. Comparing the total number of feasible partitions with the total number of possible partitions, counted by Bell numbers, we get approximate densities of $100, 60, 11.82, 0.23, 3.33\times 10^{-4}, 2.23\times 10^{-8}, 5.31\times 10^{-14}, 3.40\times 10^{-21}, 4.64\times 10^{-30}, 1.09\times 10^{-40}$, and 3.68×10^{-53} percent respectively. With a decrease in density in the order of 10^{10} and the corresponding increase in computing time to run the sorting algorithms, and another decrease in density in the order of 10^{13} going from K_{11} to K_{12}, we chose not to run the sorting algorithm

for K_{12}.

Our code is available to the reader upon request.

References

[1] M. Axenovich, P. Iverson, "Edge-colorings avoiding rainbow and monochromatic subgraphs", *Discrete Math*, vol. 308, no. 20, pp. 4710–4723, 2008.

[2] M. Axenovich, R. Jamison, "Canonical Pattern Ramsey numbers", *Graphs and Combinatorics*, vol 21, no. 2, pp 145-160, 2005.

[3] F. R. K. Chung, R. L. Graham, "Edge-colored complete graphs with percisely colored subgraphs", *Combinatorica*, vol. 3, no. 3, pp. 315–324, 1983.

[4] P. Erdős, M. Simonovits, V.T. Sós, "Anti-Ramsey theorems," in *Infinite and finite sets, Colloq., Keszthely (1973), Vol. II. Colloq. Math. Soc. Janos Bolyai, Vol. 10,* Amsterdam, North-Holland, 1975, pp. 633–643.

[5] S. Fujita, C. Magnant, K. Ozeki, "Rainbow generalizations of Ramsey theory: a survey," *Graphs Combin.*, vol. 26, no. 1, pp. 1–30, 2010.

[6] A. Goughe, D. Hoffman, P. Johnson, L. Nunley, L. Paben, "Edge-colorings of K_n Which Forbid Rainbow Cycles," *Utilitas Mathematica*, vol. 83, no.2, pp. 219–232, 2010.

[7] R. L. Graham, D. E. Knuth, O. Patashnik, *Concrete Mathematics A Foundation for Computer Science (2nd Edition)*, Reading, MA: Addison-Wesley, 1994. ISBN-13: 978-0201558029.

[8] T. Jiang, D. Mubayi, Z. Tuza, V. Voloshin, D. B. West, "The chromatic spectrum of mixed hypergraphs", *Graphs Combin.*, vol. 18, no. 2, pp. 309–318, 2002.

[9] H. Lefmann, V. Rödl, R. Thomas, "Monochromatic vs multicolored paths," *Graphs Combin.*, vol. 8, no. 3, pp. 323–332, 1992.

[10] D. Kral, "Mixed hypergraphs and other coloring problems," *Discrete Math.*, vol. 307, no. 7-8, pp. 923–938, 2007.

[11] M. Orlov, "Efficient Generation of Set Partitions," Engineering and Computer Sciences, University of Ulm, Holland, Tech. Rep., 2002.

[12] S. Sen, *New Systems and Algorithms for Scalable Fault Tolerance*, Ph.D. dissertation, Princeton University, 2013, 154 p.

[13] D. Slutzky, "Recursive Formulae for the Chromatic Polynomials of Complete r-uniform Mixed Interval Hypergraphs," *Ars Combinatoria*, to be published.

[14] V. Voloshin, "The mixed hypergraphs," *Computer Science Journal of Moldova*, vol. 1, no. 1, pp. 45–52, 1993.

[15] V. Voloshin, "On the upper chromatic number of a hypergraph," *Australas. J. Combin.,* vol. 11, pp. 25–45, 1995.

[16] V. Voloshin, *Coloring Mixed Hypergraphs: Theory, Algorithms and Applications*, American Mathematical Society, 2002. ISBN-13: 978-0821828120.

David Slutzky, [1] Vitaly Voloshin [2]

Received May 8, 2016

[1] University of North Georgia, Watkinsville, Georgia, USA
E–mail: david.slutzky@ung.edu
[2] Troy University, Troy, Alabama, USA
E–mail: vvoloshin@troy.edu

Computer Science Journal of Moldova, vol.**24**, no.2(71), 2016

Bell Numbers of Complete Multipartite Graphs

Julian Allagan & Christopher Serkan

Abstract

The *Stirling number* $S(G;k)$ is the number of partitions of the vertices of a graph G into k nonempty independent sets and the number of all partitions of G is its *Bell number*, $B(G)$. We find $S(G;k)$ and $B(G)$ when G is any complete multipartite graph, giving the upper bounds of these parameters for any graph.

Keywords: Bell number, Bell polynomial, Partition, Stirling numbers.

1 Introduction

Throughout this paper, the graph $G = (V, E)$ will be a finite simple graph with vertex set $V = V(G)$ and edge set $E = E(G)$. The *join* of two graphs G_1 and G_2, denoted by $G_1 \vee G_2$, is the graph G whose vertex set is $V(G) = V(G_1) \cup V(G_2)$, a disjoint union, and whose edge set is $E(G) = E(G_1) \cup E(G_2) \cup \{u_1u_2 \mid u_1 \in V(G_1),\ u_2 \in V(G_2)\}$. For example, $\overline{K}_{n_1} \vee \overline{K}_{n_2} \vee \ldots \vee \overline{K}_{n_l} = K(n_1, n_2, \ldots, n_l)$, a complete l-partite graph ($l \geq 1$) with parts sizes $n_1, n_2, \ldots, n_l$. The special case when $l = 1$, $G = \overline{K}_{n_1} = E_{n_1}$, the null graph. See Figure 1 for the case when $l = 2$ and $n_1 = n_2 = 3$. A *partition* $\sigma = \sigma(n)$ of an n-set X is a set of nonempty subsets of X such that each element of X is in exactly one of the subsets of X. The elements or parts of σ are often called *blocks*, and the number of blocks of σ is its *rank*. For simplicity, we refer to a partition of rank k as a k-partition. B. Duncan and R. Peele [5] called the number of k-partitions of a graph G the *(graphical) Stirling number* of G and it is denoted by $S(G, k)$; this is the number of (vertex) independent sets of G. So when $G = E_n$, $S(G;k) = \left\{ {n \atop k} \right\}$, the Stirling number of the second kind, which counts the number of

k-partitions of a set of n elements. The (total) number of distinct partitions of G is its *Bell number* which we denote by $B(G)$. In other words, $B(G) = \sum_{k=1}^{n} S(G;k)$ and when $G = E_n$, $B(E_n) = \sum_{i=1}^{n} \left\{ {n \atop i} \right\} = B_n$, where B_n is the nth Bell number. It is well documented that the exponential generating function for Bell numbers is $\exp(e^x - 1)$ i.e., $\sum_{n\geq 0} \frac{B_n}{n!} x^n = e^{e^x - 1}$. We call the rank-generating function $F(G;x) = \sum_{k=1}^{n} S(G;k)x^k$ the *partition polynomial* of the graph G. Some basic properties of this polynomial were first studied by Korfhage [9] and later by Brenti et al. [2], [3], [17]. D. Galvin and D.T. Thanh have recently named this polynomial, the Stirling polynomial [6]. It is worth noting that $F(G;1) = B(G)$. If $S(G;k) = \left\{ {n \atop k} \right\}$, then $F(G;x) = F(x)$ is the Bell polynomial which is a very well studied mathematical tool in combinatorial analysis [4], [13]. When x^i is replaced by the falling factorial $x^{\underline{i}} = x(x-1)(x-2)\ldots(x-i+1)$, the polynomial $F(G;x) = \chi(G;x) = \sum_{k=1}^{n} S(G;k)x^{\underline{k}}$ is the chromatic polynomial, which gives the number of proper colorings of a graph with n vertices, using at most x colors (see for e.g., [1], [12], [14], [15]). Clearly, there are $\chi(E_n;x) = \sum_{k=1}^{n} \left\{ {n \atop k} \right\} x^{\underline{k}} = x^n$ colorings (with no restriction) of E_n. We call the sequence $\Big(S(G;p)\Big)_{c\leq p\leq |V(G)|}$ the *partition sequence*. When $c = \chi$, the chromatic number, this sequence is referred to as chromatic vector by Goldman et al. [7] and as chromatic spectrum by Voloshin [16].

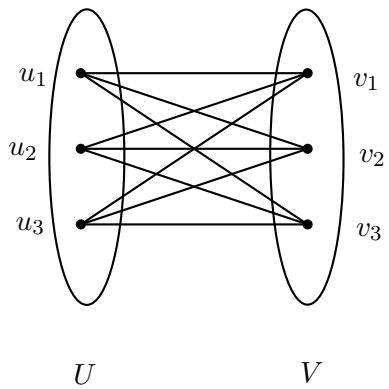

Figure 1: Complete Bipartite $K(3,3)$

The Bell numbers of special graphs have been well researched [4]–[6], [8], [10], [18]. Recently, W. Yan [18] showed that the Bell number of a k-tree on n vertices is B_{n-k}, $k \geq 1$; this is the number of k-nonconsecutive partitions of a set with n elements. This result is a generalization of that of A. O. Munagi [11] for paths, and the work of Duncan and Peele [5] for generalized paths and acyclic graphs. We record these and a few other results for some special graphs G in Table 1; the values found in the table can be obtained through either of the following:

(a) The transformation $S(G;k) = \dfrac{1}{k!}\sum_{x\geq 0}\binom{k}{x}(-1)^{k-x}\chi(G;x)$

(b) The recursion $F(G;x) = F(G-e;x) - F(G/e;x)$, where $G-e$ and G/e are the deletion and contraction graph operations on the edge e in G, respectively.

This paper was primarily inspired by the work of the previously mentioned authors as it adds to the results listed in Table 1, by extending those in rows 1, 2 and 6. In Section 2 we give a basic example of a general case which we present in Section 3.

In Table 1, $E_n, K_n, T_n^m, S_n^{(m)}, P_n^{(m)}, C_n, W_n, K(m,n)$, and $\overline{G}$ denote a null graph, a complete graph, an m-tree, an m-star, an m-path, a cycle, a wheel, a complete bipartite graph, and the complement of G, respectively. For basic notions of these graphs see [19]. This table is adapted from the one produced by Z. Kereskényi-Balogh and G. Nyul [6].

Table 1. Partition sequences and Bell numbers of some graphs

G	**Partition sequence**	$B(G)$
E_n	$\left(\left\{{n \atop p}\right\}\right)_{p\geq 1}$	B_n
K_n	$\left(1\right)_{p=n}$	1
$T_n^m, S_n^{(m)}, P_n^{(m)}$	$\left(\left\{{n-m \atop p-m}\right\}\right)_{p\geq m}$	B_{n-m}
C_n	$\left(\sum_{j=p-1}^{n-1}(-1)^{n-1-j}\left\{{j \atop p-1}\right\}\right)_{p>2}$	$\sum_{j=1}^{n-1}(-1)^{n-1-j}B_j$
W_n, $n\geq 4$	$\left(\sum_{j=p-2}^{n-2}(-1)^{n-2-j}\left\{{j \atop p-1}\right\}\right)_{p>3}$	$\sum_{j=1}^{n-2}(-1)^{n-2-j}B_j$
$K(m,n)$	$\left(\sum_{j=1}^{p}\left\{{m \atop j}\right\}\left\{{n \atop p-j}\right\}\right)_{p>2}$	$B_m\cdot B_n$
$\overline{S_n}$, $n\geq 2$	$\left(n-1\right)_{p=n-1}$ and $\left(1\right)_{p=n}$	n
$\overline{P_n}$, $n\geq 2$	$\left(\binom{p}{n-p}\right)_{p\geq\lceil\frac{n}{2}\rceil}$	F_{n+1} (Fibonacci number)
$\overline{C_n}$, $n\geq 4$	$\left(\frac{n}{p}\binom{p}{n-p}\right)_{p\geq\lceil\frac{n}{2}\rceil}$	L_n (Lucas number)

2 Example

Consider the bipartite graph $G = K(3,3)$ with parts $U = \{u_1, u_2, u_3\}$ and $V = \{v_1, v_2, v_3\}$ in Figure 1. Because $U = V = E_3$, it follows that each set has the partition sequence $(\left\{{3 \atop 1}\right\}, \left\{{3 \atop 2}\right\}, \left\{{3 \atop 3}\right\})$ which is $(1,3,1)$ and the distinct partitions of, say U are:

- rank 1: $u_1u_2u_3$
- rank 2: $u_1|u_2u_3$; $u_2|u_1u_3$; $u_3|u_1u_2$
- rank 3: $u_1|u_2|u_3$

Hence the partition polynomial $F(U;x) = F(V;x) = 1x^1 + 3x^2 + 1x^3$. Since no element $x \in U$ can be in the same block as an element $y \in V$, a q-partition of G is therefore composed of all the i-partitions of U and all the j-partitions of V such that $i + j = q$. If we denote by a_i and b_j the terms of the partition sequences of U and V respectively, for each $1 \leq i, j \leq 3$, then the q-partitions of G form the 3×3 array (Table 2).

Table 2. An array of q-partitions of $G = K(3,3)$ with $q = 2, \ldots, 6$

	b_1	b_2	b_3
a_1	a_1b_1	a_1b_2	a_1b_3
a_2	a_2b_1	a_2b_2	a_2b_3
a_3	a_3b_1	a_3b_2	a_3b_3

Observe that the indices of the off-diagonal entries add up to q, the power of x in the polynomial $F(G;x) = a_1b_1x^2 + (a_1b_2 + a_2b_1)x^3 + (a_1b_3 + a_2b_2 + a_3b_1)x^4 + (a_2b_3 + a_3b_2)x^5 + a_3b_3x^6$. Since $a_1 = a_3 = b_1 = b_3 = 1$ and $a_2 = b_2 = 3$, it follows that $F(G;x) = 1x^2 + 6x^3 + 11x^4 + 6x^5 + 1x^6 = F(U;x) \cdot F(V;x)$. The corresponding partition sequence is $(0,1,6,11,6,1)$ with $S(G;1) = 0$, $S(G;2) = 1$, $S(G;3) = 6$, $S(G;4) = 11$, $S(G;5) = 6$, $S(G;6) = 1$ and the Bell number is $B(G) = 25 = B(V) \cdot B(U)$.

3 Bell numbers of complete multipartite graphs

Theorem 1. *Suppose $G_1, \ldots, G_l$ are graphs, each with a partition vector $(a_k^1, \ldots, a_k^{n_k})$, $1 \leq k \leq l$. If $G = G_1 \vee \ldots \vee G_l$, then the partition polynomial*

$$F(G;x) = \sum_{q=l}^{n_1+\ldots+n_l} \Big(\sum_{\substack{(j_1,\ldots,j_l) \\ j_1+\ldots+j_l=q}} a_1^{j_1} \ldots a_l^{j_l} \Big) x^q \text{ for all } l \geq 1.$$

Proof. When $l = 1$, $F(G;x) = \sum_{\substack{q=1 \\ j_1=q}}^{n_1} a^{j_1} x^q$ is the partition polynomial of $G = G_1$. For $1 \leq k \leq l$, $F(G_k;x) = \sum_{i=1}^{n_k} a_k^i x^i$, by definition. Now suppose each of the following k columns represents the terms of each partition polynomial, $F(G_k;x)$.

$$\begin{array}{cccc} a_1^1 x & a_2^1 x & \ldots & a_l^1 x^1 \\ a_1^2 x^2 & a_2^2 x^2 & \ldots & a_l^2 x^2 \\ \vdots & \vdots & & \vdots \\ a_1^{n_1} x^{n_1} & a_2^{n_2} x^{n_2} & \ldots & a_l^{n_l} x^{n_l} \end{array} \tag{1}$$

Since, for any partition of $V(G)$, no element $u \in V(G_i)$ can be in the same block with an element $w \in V(G_j)$, $i \neq j$, this implies that $F(G;x) = \prod_{k=1}^{l} F(G_k;x)$. Moreover, a term of $F(G;x)$ that involves, say x^q, is obtained by taking $a_k^{j_k} x^{j_k}$ from the kth column and forming the product $\prod_{k=1}^{l} a_k^{j_k} x^{j_k}$, with the exponents of x satisfying $\sum_{k=1}^{l} j_k = q$. This implies that all the terms of x^q are $\sum_{\substack{(j_1,\ldots,j_l) \\ \sum j_k = n}} \prod_{k=1}^{l} a_k^{j_k} x^{j_k}$. Because a sum over all the terms of x^q for $l \leq q \leq \sum_{i=1}^{n_k}$ is the polynomial $F(G;x)$, this gives the result.

□

Corollary 1. *The partition polynomial of a complete l-partite graph with part sizes n_i is*

$$F(G;x) = \sum_{q=l}^{n_1+\ldots+n_l} \Big(\sum_{\substack{(j_1,\ldots,j_l) \\ j_1+\ldots+j_l=q}} \left\{ {n_1 \atop j_1} \right\} \ldots \left\{ {n_l \atop j_l} \right\} \Big) x^q.$$

Proof. Because $G = \overline{K}_{n_1} \vee \overline{K}_{n_2} \vee \ldots \vee \overline{K}_{n_l}$ and $a_k^j = \left\{ {n_k \atop j} \right\}$ for $1 \leq j, k \leq l$, the result follows from Theorem 1. □

Corollary 2. *The partition sequence of a complete l-partite graph with part sizes $n_i \geq 1$ is* $\Big(\sum_{\substack{(j_1,\ldots,j_l) \\ \sum j_k=p}} \left\{ {n_1 \atop j_1} \right\} \ldots \left\{ {n_l \atop j_l} \right\} \Big)_{p \geq l}$, $l \geq 1$.

Since each $F(\overline{K}_{n_k};1) = B_{n_k}$ and $F(G;1) = \prod_{k=1}^{l} F(\overline{K}_{n_k};1)$, the next result follows.

Corollary 3. *The Bell number of a complete l-partite graph with parts sizes n_i is* $B(G) = \prod_{i=1}^{l} B_{n_i}$.

Remarks. Observe that Corollaries 2 and 3 generalize the result on row 6 of Table 1, which extends those of rows 1 and 2. As mentioned in the introduction, the lower bound for the Stirling numbers of any l-colorable graph H is its chromatic number $\chi(H) = l$. Because every l-colorable graph H is a subgraph of some complete l-partite graph G, the previous two results give the upper bounds for $S(H;k)$ and $B(H)$.

References

[1] C.D. Birkhoff and D.C. Lewis, "Chromatic polynomials," *Trans. Amer. Math. Soc.*, vol. 60, pp. 335–351, 1946.

[2] F. Brenti, "Expansions of chromatic polynomials and log-concavity," *Trans. Amer. Math. Soc.*, vol. 332, no2, pp. 729–756, August 1992.

[3] F. Brenti, G. Royle and D. Wagner, "Location of zeros of chromatic and related polynomials of graphs," *Canad. J. Math.*, vol. 46, pp. 55–80, 1994.

[4] C.B. Collins, "The role of Bell polynomials in Integration," *J. Comput. Appl. Math.*, vol. 131, pp. 195–222, 2001.

[5] B. Duncan and R. Peele, "Bell and Stirling numbers for graphs," *J. Integer Seq.*, vol. 12, Article 09.7.1, 2009.

[6] D. Galvin and D.T. Thanh, "Stirling numbers of forests and cycles," *Electron. J. Combin.*, vol. 20, no. 1, P73, 2013.

[7] J. Goldman, J. Joichi and D. White, "Rook Theory III. Rook polynomials and the Chromatic structure of graphs," *J. Combin. Theory Ser. B*, vol. 25, pp. 135–142, 1978.

[8] Z. Kereskényi-Balogh and G. Nyul, "Stirling numbers of the second kind and Bell numbers for graphs" *Australas. J. of Combin.*, vol. 58, no.2, pp. 264–274, 2014.

[9] R. Korfhage, "σ-polynomials and graph coloring," *J. Combin. Theory Ser. B*, vol. 24, pp. 137–153, 1978.

[10] A. Mohr and T.D. Porter, "Applications of chromatic polynomials involving Stirling numbers," *J. Combin. Math. Combin. Comput.*, vol. 70, pp. 57–64, 2009.

[11] A.O. Munagi, "k-complementing subsets of nonnegative integers," *Int. J. Math. Math. Sci.*, pp. 215–224, 2005.

[12] R.C. Read, "An introduction to chromatic polynomials," *J. Combin. Theory*, vol. 4, pp. 52–71, 1968.

[13] J. Riordan, *An Introduction to Combinatorial Analysis,* New York, NY, USA: Wiley, 1958.

[14] R.P. Stanley, "Acyclic orientations of graphs," *Discrete Math.*, vol. 5, pp. 171–178, 1973.

[15] W.T. Tutte, "On chromatic polynomials and the golden ratio," *J. Combin. Theory*, vol. 9, pp. 289–296, 1970.

[16] V.I. Voloshin, *Coloring Mixed Hypergraphs: Theory, Algorithms and Applications* (AMS and Fields Institute Monographs), Providence, RI, USA: AMS, 2002.

[17] D. Wagner, "The partition polynomial of a finite set system," *J. Combin. Theory Ser. A*, vol. 56, pp. 138–159, 1991.

[18] W. Yang, "Bell numbers and k-trees," *Discrete Math.*, vol. 156, pp. 247–252, 1996.

[19] D. West, *Introduction to Graph Theory*, Prentice Hall, 2001.

Julian Allagan, Christopher Serkan

Received June 9, 2016

Julian Allagan
University of North Georgia
Watkinsville, Georgia, U.S.A
E–mail: `julian.allagan@ung.edu`

Christopher Serkan
University of North Georgia
Watkinsville, Georgia, U.S.A
E–mail: `christopher.serkan@ung.edu`

Computer Science Journal of Moldova, vol.**24**, no.2(71), 2016

Hat problem on graphs with exactly three cycles

Tayebe Balegh, Nader Jafari Rad

Abstract

This paper is devoted to investigation of the hat problem on graphs with exactly three cycles. In the hat problem, each of n players is randomly fitted with a blue or red hat. Everybody can try to guess simultaneously his own hat color by looking at the hat colors of the other players. The team wins if at least one player guesses his hat color correctly, and no one guesses his hat color wrong; otherwise the team loses. The aim is to maximize the probability of winning. Note that every player can see everybody excluding himself. This problem has been considered on a graph, where the vertices correspond to the players, and a player can see each player to whom he is connected by an edge. We show that the hat number of a graph with exactly three cycles is $\frac{3}{4}$ if it contains a triangle, and $\frac{1}{2}$ otherwise.

Keywords: Hat problem, Strategy.
2010 Mathematics Subject Classification: 05Cxx.

1 Introduction

In the hat problem there are n players who may coordinate a strategy before the game begins. Each player gets a hat whose color is selected randomly and independently to be blue with probability 1/2 and red otherwise. Each player can see the colors of all other hats but not of his own. Simultaneously, each player may guess a color or pass. The players win if at least one player guesses correctly the color of his own hat, and no player guesses wrong. The goal is to find a strategy that

maximizes the probability of winning. This maximum probability is called the value of the game. This problem was formulated by Ebert [2], and further considered for example in [4], [5], [11].

The hat problem on a graph was considered by Krzywkowski [6]. The players are placed on the vertices of a graph, and a player can only see the colors of hats of his neighbors. The requirement for winning remains the same. If the graph is a complete graph, this is exactly Ebert's original problem. Krzywkowski in [6] showed that if the graph is a tree, the value of the corresponding game is 1/2. In [7] the same result is shown when the graph is a cycle on four vertices. The hat problem on bipartite graphs, cycles, unicyclic graphs, and graphs with exactly two cycles are studied in [1], [3], [7]–[10], [12]. In this paper we study the hat problem in graphs with exactly three cycles. Let $h(G)$ denotes the value of the hat problem on a graph G. We shall prove the following.

Theorem 1 *Let G be a graph with exactly three cycles. Then $h(G) = \frac{3}{4}$ if G contains a triangle, and $h(G) = \frac{1}{2}$ otherwise.*

2 Notations

For notation and terminology not given here we refer to [13]. Let $G = (V(G), E(G))$ be a graph. For a vertex $v \in V(G)$, the *open neighborhood* of v, is $N_G(v) = \{x \in V(G) : vx \in E(G)\}$. The *degree* of a vertex v is $\deg_G(v) = \deg(v) = |N_G(v)|$. We say that a vertex v is *neighborhood-dominated* if there is some other vertex u such that $N_G(v) \subseteq N_G(u)$. If H is a subgraph of G, then we write $H \subseteq G$.

Let $V(G) = \{v_1, v_2, ..., v_n\}$. A function $c : V(G) \to \{b, r\}$ is a vertex coloring, where b refers to the blue color and r refers to the red color. If $v_i \in V(G)$, then $c(v_i)$ is the color of v_i. By a case for the graph G we mean a sequence $(c(v_1), c(v_2), ..., c(v_n))$. We denote the set of all cases for the graph G by $C(G)$. Note that $|C(G)| = 2^{|V(G)|}$. If $v_i \in V(G)$, then by s_i we denote a function $s_i : V(G) \to \{b, r, *\}$, where $s_i(v_j)$ is the first letter of the color of v_j if v_i sees v_j, and mark $*$ otherwise, that is, $s_i(v_j) = c(v_j)$ if $v_j \in N_G(v_i)$, while $s_i(v_j) = *$ if

$v_j \in V(G) - N_G(v_i)$. By a situation of the vertex v_i in the graph G we mean the sequence $(s_i(v_1), s_i(v_2), ..., s_i(v_n))$. The set of all possible situations of v_i in the graph G is denoted by $St_i(G)$. Observe that $|St_i(G)| = 2^{|N_G(v_i)|}$. If $v_i \in V(G)$, then we say that a case $(c_1, c_2, ..., c_n)$ for the graph G corresponds to a situation $(t_1, t_2, ..., t_n)$ of the vertex v_i in the graph G if it is created from this situation only by changing every mark $*$ to the letter b or r. So, a case corresponds to a situation of v_i if every vertex adjacent to v_i, in that case has the same color as in that situation. To every situation of the vertex v_i in the graph G correspond $2^{|V(G)|-\deg_G(v_i)}$ cases, because every situation of v_i has $|V(G)| - \deg_G(v_i)$ mark $*$.

By a statement of a vertex we mean its declaration about the color it guesses it is. By the effect of a case we mean a win or a loss. According to the definition of the hat problem, the effect of a case is a win if at least one vertex states its color correctly and no vertex states its color wrong. The effect of a case is a loss if no vertex states its color or somebody states its color wrong. By a guessing instruction for the vertex $v_i \in V(G)$ (denote by g_i) we mean a function $g_i : St_i(G) \to \{b, r, p\}$ which, for a given situation, gives the first letter of the color v_i guesses it is or a letter p if v_i passes. Thus a guessing instruction is a rule which determines the conduct of the vertex v_i in every situation. By a strategy for the graph G we mean a sequence $(g_1, g_2, ..., g_n)$. By $F(G)$ we denote the family of all strategies for the graph G.

Let $v_i \in V(G)$ and $S \in F(G)$. We say that v_i never states its color in the strategy S if v_i passes in every situation. We say that v_i always states its color in strategy S if v_i states its color in every situation, that is, for every $T \in St_i(G)$ we have $g_i(T) \in \{b, r\}$ ($g_i(T) \neq p$, equivalently). If $S \in F(G)$, then by $Cw(S)$ and $Cl(S)$ we denote the sets of cases for the graph G in which the team wins or loses, respectively. Observe that $|Cw(S)| + |Cl(S)| = |C(G)|$. Consequently, by the chance of success of the strategy S we mean the number $p(S) = \frac{|Cw(S)|}{|C(G)|}$. By the hat number of the graph G we mean the number $h(G) = max\{p(S) : S \in F(G)\}$. Note that $p(S) \leq h(G)$. We say that the strategy S is optimal for the graph G if $p(S) = h(G)$. By $F^0(G)$ we denote the family of all optimal strategies for the graph G.

3 Known results

In this section we state some known results that we need to prove our main result. We denote by P_n, C_n and K_n the path, the cycle and the complete graph with n vertices, respectively. We begin with the following theorem.

Theorem 2 (Krzywkowski, [6]) *If H is a subgraph of G, then $h(H) \leq h(G)$.*

Corollary 1 (Krzywkowski, [6]) *For every graph G, $h(G) \geq \frac{1}{2}$.*

Let $\omega(G)$ denotes the *clique number* of a graph G, i.e. the maximum number of vertices that each pair of them are adjacent. Also let $\chi(G)$ denotes the *chromatic number* of G, i.e. the minimum number of colors in a vertex coloring such that adjacent vertices receive different colors. Feige, [3] presented the following important results.

Theorem 3 (Feige, [3]) *For every graph, $h(G) = h(K_{\omega(G)})$, if $\chi(G) = \omega(G)$.*

Theorem 4 (Feige, [3]) *If $\omega(G) + 1$ is a power of 2, then $h(G) = \frac{\omega(G)}{\omega(G)+1}$.*

Lemma 1 (Feige, [3]) *If v is a neighborhood-dominated vertex of a graph G, then $h(G) = h(G - v)$.*

Lemma 2 (Feige, [3]) *If a graph G is a disjoint union of two graphs G_1 and G_2, then $h(G) = \max\{h(G_1), h(G_2)\}$.*

We denote by $G_1 \cup G_2$ the disjoint union of two graphs G_1 and G_2. The hat number of several classes of graphs including paths, cycles, unicyclic graphs, and graphs with precisely two cycles are determined as follows.

Theorem 5 (Krzywkowski, [6]) *For every path P_n we have $h(P_n) = \frac{1}{2}$.*

Theorem 6 (Feige, [3], Krzywkowski, [7], [8]) *For every cycle C_n with $n > 3$, $h(C_n) = \frac{1}{2}$.*

Lemma 3 (Krzywkowski, [10]) *If G is a unicyclic graph with no triangle, then $h(G) = \frac{1}{2}$.*

Theorem 7 (Balegh, Jafari Rad, [1]) *If G is a graph with no triangle and exactly two cycles, then $h(G) = \frac{1}{2}$.*

The next two theorems consider optimal strategies such that some vertex always (never, respectively) states its color.

Theorem 8 (Krzywkowski, [6]) *Let v be a vertex of a graph G. If $S \in F^0(G)$ is a strategy such that v always states its color, then $h(G) = \frac{1}{2}$.*

Theorem 9 (Krzywkowski, [6]) *Let v be a vertex of a graph G. If $S \in F^0(G)$ is a strategy such that v never state its color, then $h(G) = h(G - v)$.*

Remark 1 *Let the strategy S is optimal for a graph G, then we have $h(G) = p(S)$, we get $p(S) \geq \frac{1}{2}$.*

The next lemma is about the non-necessity of statements of any further vertices in a case in which some vertex already states its color.

Lemma 4 (Krzywkowski, [7]) *Let G be a graph and let S be a strategy for G. Let C be a case in which some vertex states its color. Then a statement of any other vertex cannot improve the effect of the case C.*

4 Proof of Theorem 1

Let G be a graph with exactly three cycles. Assume that $\delta(G) = 1$. Clearly any vertex of degree one is a neighborhood-dominated vertex. If y_1 is a vertex of degree one in G, then by Lemma 1, $h(G - y_1) = h(G)$.

If $\delta(G - y_1) = 1$ and y_2 is a vertex of degree one in $G - y_1$, then by Lemma 1, $h(G-y_1-y_2) = h(G-y_1) = h(G)$. Continuing this process, there is an integer k such that $h(G) = h(G - y_1 - y_2 - ... - y_k)$, and $\delta(G - y_1 - y_2 - ... - y_k) \geq 2$. Thus we may assume that $\delta(G) \geq 2$. Assume G has a triangle. Clearly $\omega(G) = 3$ since G has exactly three cycles. Then by Theorem 4, we have $h(G) = \frac{3}{4}$. Thus for the next we assume that G contains no triangle. The following lemma plays an important role for the next.

Lemma 5 *Suppose $P = v_1v_2v_3v_4$ is a path in G with $\deg_G(v_2) = \deg_G(v_3) = 2$, and $v_4 \notin N_G(v_1)$. Let H be the graph obtained from G by deleting the vertices v_2 and v_3 and, adding an edge between v_1 and v_4. Then $h(G) \leq h(H)$.*

Proof. Let H_1 be obtained from G by adding the edge v_1v_4. By Theorem 2, $h(G) \leq h(H_1)$. Then v_3 is a neighborhood dominated vertex in H_1, and thus by Theorem 2 and Lemma 1, $h(H_1) = H(H_1 - v_3)$. But v_2 is a neighborhood dominated vertex in $H_1 - v_3$, and thus by Theorem 2 and Lemma 1, $H(H_1 - v_3) = h(H_1 - v_3 - v_1)$. Now $h(G) \leq h(H_1) = h(H_1 - v_3) = h(H_1 - v_3 - v_1) = h(H)$, as desired. $\square$

4.1 G has no cut-vertex

Since G has no cut-vertex, it is obtained from a cycle by adding a path $P = x_0x_1....x_k$ between two non-consecutive vertices u and v, where $u = x_0$ and $v = x_k$. Thus G contains two cycles C_1 and C_2 such that $V(C_1) \cap V(C_2) = \{x_0, ..., x_k\}$. Let $|V(C_1)| = n_1$ and $|V(C_2)| = n_2$. If both n_1 and n_2 are even, then $\chi(G) = \omega(G) = 2$, and so by Theorem 3, $h(G) = \frac{1}{2}$. Thus assume that at least one of n_1 or n_2 is odd. We aim to obtain a graph G^* with $h(G) \leq h(G^*)$ and $h(G^*) = 1/2$, and then the result follows by Theorem 1. We do this in some stages, and in each stage of the proof, without loss of generality, we assume that in each stage G has the properties of the desired G^*.

By applying Lemma 5, we may assume that $k \leq 3$.

Lemma 6 *If $k = 1$, then $h(G) \leq \frac{1}{2}$.*

Proof. By applying Lemma 5, we may assume that $n_2 = 5$ and $n_1 \in \{4, 5\}$. Assume first that $n_1 = 5$. Let $n_1 = n_2 = 5$, $C_1 = x_0x_1a_1a_2a_3x_0$, and $C_2 = x_0x_1b_1b_2b_3x_0$. Let $G_1 = G + b_1a_2$. Then a_1 is a neighborhood dominated vertex in G_1, and thus by Theorem 2 and Lemma 1, $h(G) \leq h(G_1) = h(G_1 - a_1)$. Let $G_2 = G_1 - a_1$, and $G_3 = G_2 + a_2b_3$. Then b_2 is a neighborhood dominated vertex in G_3, and thus by Theorem 2 and Lemma 1, $h(G_3) = h(G_3 - b_2)$. Let $G_4 = G_3 - b_2$. Then a_3 is a neighborhood dominated vertex in G_4, and thus by Theorem 2 and Lemma 1, $h(G_4) = h(G_4 - a_3)$. But $G_4 - a_3$ is a cycle, and by Theorem 6, $h(G_4 - a_3) = 1/2$. Thus $h(G) \leq h(G_1) \leq h(G_2) \leq h(G_3) \leq h(G_4) \leq 1/2$.

Next assume that $n_1 = 4$. Let $C_1 = abx_1x_0a$, where $N_G(b) = \{a, x_1\}$. Since $N_G(b) \subseteq N_G(x_0)$, by Lemma 1, $h(G) = h(G - b)$. But $G - b$ is a unicyclic graph, and so by Lemma 3, $h(G) = h(G - b) = 1/2$. □

Lemma 7 *If $k = 2$, then $h(G) \leq 1/2$.*

Proof. Assume that $k = 2$. By applying Lemma 5, we may assume that $n_2 = 5$, and $n_1 \in \{4, 5\}$. First assume that $n_1 = 5$. Let $C_1 = x_0x_1x_2a_1a_2x_0$, and $C_2 = x_0x_1x_2b_1b_2x_0$. Let $G_1 = G + b_2a_1$. Then b_1 is a neighborhood dominated vertex in G_1, and thus by Theorem 2 and Lemma 1, $h(G) \leq h(G_1) = h(G_1 - b_1)$. Let $G_2 = G_1 - b_1$. Then b_2 is a neighborhood dominated vertex in G_2, and thus by Theorem 2 and Lemma 1, $h(G_2) = h(G_2 - b_2)$. But $G_2 - b_2$ is a cycle, and by Theorem 6, $h(G_2 - b_2) = 1/2$. Thus $h(G) \leq h(G_1) \leq h(G_2) \leq 1/2$.

Next assume that $n_1 = 4$. Then C_1 has a neighborhood-dominated vertex, say x, which $x \notin \{x_0, x_1, x_2\}$, and thus by Theorem 2 and Lemma 1, we find that $h(G) \leq h(G - x) = h(C_2) = \frac{1}{2}$, implying that $h(G) \leq 1/2$. □

Lemma 8 *If $k = 3$, then $h(G) \leq 1/2$.*

Proof. Assume that $k = 3$. Since G has no triangle, $\{n_1, n_2\} \neq \{4, 5\}$. If $n_1 = 4$, then x_1 is a neighborhood-dominated vertex, and thus by Theorem 2 and Lemma 1, we find that $h(G) \leq h(G - x_1) = 1/2$. Thus

$n_1 > 4$, and similarly $n_2 > 4$. Let n_1 be even. Let $x_1x_0v_1v_2$ be a path on C_1 with $v_1 \neq x_1$, and let H be obtained from G by joining x_1 to v_2. Then v_1 is a neighborhood-dominated vertex in H, and thus by Lemma 1, $h(G) \leq h(H - v_1)$. But $h(H - v_1) \leq 1/2$ by Lemma 7. Thus $h(G) \leq 1/2$. Similarly if n_1 is odd, then $h(G) \leq 1/2$. $\square$

4.2 G has some cut-vertex

Assume that G has precisely one cut-vertex. Then G contains precisely three cycles C_1, C_2 and C_3 with one common vertex, say w. For convenience we denote G by $G_1(n_1, n_2, n_3)$, where $n_i = |V(C_i)|$ for $i = 1, 2, 3$. If n_i is even for all $i = 1, 2, 3$, then by Theorem 3, we have $\chi(G) = \omega(G) = 2$, and so $h(G) = \frac{1}{2}$. Thus without loss of generality assume that n_1 is odd. By applying Lemma 5, we may assume that $n_1 = 5$, $n_2 \in \{4, 5\}$ and $n_3 \in \{4, 5\}$. Assume that $n_2 = 4$. Let $V(C_2) = \{a, b, c, w\}$, where $N_G(b) = \{a, c\}$. Then b is a neighborhood-dominated vertex, and thus by Theorem 2 and Lemma 1, we find that $h(G) = h(G - b)$. But $G - b$ is a graph with exactly two cycles, and by Theorem 7, $h(G) = 1/2$. Thus we assume that $n_2 = n_3 = 5$. Thus $G = G_1(5, 5, 5)$.

Let $V(G) = \{a_1, a_2, a_3, a_4, v, b_1, b_2, b_3, b_4, c_1, c_2, c_3, c_4\}$, where $N(v) = \{a_1, a_4, b_1, b_4\}$, a_i is adjacent to a_{i+1} for $i = 1, 2, 3$, b_j is adjacent to b_{j+1} for $j = 1, 2, 3$, and c_k is adjacent to c_{k+1} for $k = 1, 2, 3$. Let $H_1 = G + a_4b_3$. Then b_4 is a neighborhood-dominated vertex in H_1, and by Theorem 2 and Lemma 1, $h(H_1) = h(H_1 - b_4)$. Let $H_2 = H_1 - b_4$ and $H_3 = H_2 + a_1b_2$. Then b_1 is a neighborhood-dominated vertex in H_3, and by Theorem 2 and Lemma 1, $h(H_3) = h(H_3 - b_1)$. Let $H_4 = H_3 - b_1$ and $H_5 = H_4 + a_3b_2$. Then a_2 is a neighborhood-dominated vertex in H_5, and by Theorem 2 and Lemma 1, $h(H_5) = h(H_5 - a_2)$. Let $H_6 = H_5 - a_2$. We now see that b_3 is a neighborhood-dominated vertex in H_6, and by Theorem 2 and Lemma 1, $h(H_6) = h(H_6 - b_3)$. But $H_6 - b_3$ is a graph with two cycles, and thus by Theorem 7 $h(H_6 - b_3) = 1/2$.

Thus

$$\begin{aligned} h(G) \leq h(H_1) &\leq h(H_2) \leq h(H_3) \\ &\leq h(H_4) \leq h(H_5) \\ &\leq h(H_6) \leq h(H_6 - b_3) = 1/2 \end{aligned}$$

as desired.

Assume now that G has at least two cut-vertices. Assume that G has two cut vertices w_1, w_2 such that $w_1 \in V(C_1)$, $w_2 \in V(C_2)$ and the shortest path from w_1 to w_2 (say P) does not intersect C_3. Let $z_1 \in N(w_2)$ be a vertex on P. Let $v_1v_2w_2z_1$ be a path on C_3, and let $H = G + v_1z_1$. Clearly by Theorem 2, we have $h(G) \leq h(H)$. Observe that v_2 is a dominated vertex. By Lemma 1, we get $h(H) = h(H - v_2)$. If $z_1 \neq w_1$, then we consider a vertex $z_2 \in N(z_1)$ on P, and continue this process. Continuing this process, we obtain a graph H^* with precisely three cycles C_1, C_3 and C_2', where $V(C_1) \cap V(C_2') = \{w_1\}$. A similar argument holds for C_1, C_3, or C_2, C_3. Thus we may assume that G has two cut vertices w_1, w_2 such that $V(C_1) \cap V(C_2) = \{w_1\}$ and $V(C_2) \cap V(C_3) = \{w_2\}$, and $w_1 \notin N(w2)$. As before, we may assume that $|V(C_i)| = n_i$ for $i = 1, 2, 3$. Also for convenience, we denote $G = G_2(n_1, n_2, n_3)$. By applying Lemma 5, we may assume that $n_1, n_3 \in \{4, 5\}$. Assume that $n_1 = 4$. Let $V(C_1) = \{a, b, c, w_1\}$, where $N_G(b) = \{a, c\}$. Since $N_G(b) \subseteq N_G(w_1)$, by Lemma 1, $h(G) = h(G - b)$. Since $G - u$ is a graph with precisely two cycles, by Theorem 7, $h(G) \leq 1/2$. Thus $n_1 = 5$ and similarly $n_2 = 5$. Assume that $n_2 \geq 4$ is even. By applying Lemma 5, we may assume that $n_2 = 4$. Let $V(C_2) = \{w_1, v_1, w_2, v_2\}$, where $w_1 \in V(C_1 \cap C_2)$ and $w_2 \in V(C_2 \cap C_3)$. Without loss of generality, observe $N_G(v_1) = \{w_1, w_2\}$. Clearly b is a neighborhood-dominated vertex, and so by Lemma 1, $h(G) = h(G - b)$. But $G - b$ is a graph with exactly two cycles, and by Theorem 7, $h(G) = 1/2$. Thus assume that $n_2 \geq 5$ is odd. By applying Lemma 5, we may assume that $n_2 = 5$.

Lemma 9 $h(G_2(5, 5, 5) = 1/2$.

Proof. Let S be an optimal strategy for G. Let us assume that some vertices, say v_i, never states its color. Then by Theorem 9, we have

$h(G) = h(G - v_i)$. If $deg(v_i) = 2$, then $G - v_i$ is a graph with precisely two cycles. By Theorem 7, we get $h(G) = h(G - v_i) = \frac{1}{2}$. If $deg(v_i) > 2$, then $G - v_i = P_4 \cup G'$, where G' is a unicyclic graph. Then by Theorems 2, 5 and Lemma 3, we get $h(G) = h(G - v_i) = max\{h(P_4), h(G')\} = \frac{1}{2}$. Thus we assume that every vertex guesses its color. If there exists a vertex that always states its color, then by Theorem 8, $h(G) = \frac{1}{2}$. Thus assume that no vertex in G always states its color. Now let us assume that every vertex states its color in at least one situation. We consider the following two possibilities.

(1) Every vertex states its color in exactly one situation.

Every statement of every vertex in any situation is wrong in exactly $2^{|V(G)|-d_G(v_i)-1}$ cases, because every situation of any vertex v_i is corresponded to $2^{|V(G)|-|N_G(v_i)|}$ cases, and in half of them the vertex v_i has the color it states it. Since every vertex states its color in exactly one situation, there are exactly 2^{12} correct statements, and then the team can win in at most 2^{12} cases, even if every of the 2^{12} correct statements is in another cases. This implies that $p(S) = \frac{|Cw(S)|}{|C(G)|} \leq \frac{1}{2}$. Since $S \in F^0(G)$, we have $h(G) \leq \frac{1}{2}$. Since by Corollary 1, we have $h(G) \geq \frac{1}{2}$, we get $h(G) = \frac{1}{2}$.

(2) There is a vertex that states its color in more than one situation.

Since we seek minimal number of cases with wrong statements, let us assume that there is a vertex, say v_i, that states its color in exactly two situations. This vertex states its color when views an even number of blue or red colors. Without loss of generality, let v_i and v_j state their colors if it view an even number of blue colors. Let S' be an optimal strategy different from S such that for any pair of vertices v_i and v_j, one of v_i or v_j states its color when views an even number of blue colors, and the other one does not state its color when views an even number of blue colors. Let v_j does not state its color when views an even number of blue colors. Then clearly the other vertex, v_i, states its color when views an even number of blue colors. By Lemma 4 the statement

of v_j cannot improve the result of any of these cases. Therefore, $p(S') \leq p(S)$. Since $S' \in F^0(G)$, then strategy S is also optimal for G. Note that if v_j never states its color in the strategy S', then $S' = S$ and we have a possibility already considered.

□

References

[1] T. Balegh, N. Jafari Rad, "Hat problem on bicyclic graphs," *Utilitas Mathematica*, to be published.

[2] T.T. Ebert, "Applications of recursive operators to randomness and complexity," Ph.D. Dissertation, University of California at Santa Barbara, 1998.

[3] U. Feige, "On optimal strategies for a hat game on graphs," *SIAM J. Discrete Math. (SIAMDM)*, vol. 24, no. 3, pp. 782–791, 2010.

[4] W. Guo, S. Kasala, M. Rao, and B. Tucker, "The hat problem and some variations," in *Advances in distribution theory, order statistics, and inference*, Stat. Ind. Technol., Birkhauser Boston, Boston, MA, 2006, pp. 459–479.

[5] R. Hod and M. Krzywkowski, "A construction for the hat problem on a directed graph," *Electronic Journal of Combinatorics*, vol. 19, pp. 30–30, 2012.

[6] M. Krzywkowski, "Hat problem on a graph," *Mathematica Pannonica*, vol. 21, pp. 3–21, 2010.

[7] M. Krzywkowski, "Hat problem on the cycle C_4," *International Mathematical Forum*, vol. 5, pp. 205–212, 2010.

[8] M. Krzywkowski, "Hat problem on odd cycles," *Houston Journal of Mathematics*, vol. 37 pp. 1063–1069, 2011.

[9] M. Krzywkowski, "The hat problem on cycles on at least nine vertices," *Ars Combinatoria*, vol. 101, pp. 3–13, 2011.

[10] M. Krzywkowski, "On the hat problem of a graph," *Opuscula Mathematica*, vol. 32, pp. 285–296, 2012.

[11] M. Krzywkowski, "On the hat problem, its variations, and their applications," *Annales Universitatis Paedagogicae Cracoviensis Studia Mathematica*, vol. 9, pp. 55–67, 2010.

[12] M. Krzywkowski, "On the hat problem on the cycle C_7," *International Journal of Contemporary Mathematical Sciences*, vol. 5, pp. 2137–2148, 2010.

[13] D.B. West, *Introduction to Graph Theory*, Prentice Hall, Inc., Upper Saddle River, NJ, 1996, xvi+512 p. ISBN: 0-13-227828-6.

Tayebe Balegh, Nader Jafari Rad

Received May 23, 2016

Department of Mathematics, Shahrood University of Technology,
Shahrood, Iran
E–mail: n.jafarirad@gmail.com

Computer Science Journal of Moldova, vol.**24**, no.2(71), 2016

Numerical solutions of Kendall and Pollaczek-Khintchin equations for exhaustive polling systems with semi-Markov delays

Gheorghe Mishkoy, Diana Bejenari, Lilia Mitev, Ionela R. Ticu

Abstract

Some analytical results for exhaustive polling systems with semi-Markov delays, such as Pollaczek-Khintchin virtual and steady state analog are presented. Numerical solutions for k-busy period, probability of states and queue length distribution are obtained. Numerical examples are presented.

Keywords: Polling systems with semi-Markov delays, Pollaczek-Khintchin formula, Kendall equation, k-busy period, probability of states, queue length, numerical algorithms.

1 Introduction

It is known that wireless networks have developed rapidly last years. For planning regional wireless networks, models and research methods of polling systems are used [1]. A polling model is a system of multiple queues accessed by a single server in a given order. Among important characteristics of these systems are the k-busy period, probability of states and queueing length [2]. We consider a queueing system of polling type with semi-Markov delays. Handling mechanism for this system is given by polling table $f : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, r\}$, where the function shows that at the stage j, $j = \overline{1, n}$, user number k, $k = \overline{1, r}$, $r \leq n$ is served (more details see in [1]). The items (messages) of the user k, according to Poisson distribution with parameter λ_k arrive. The service time for the items of class k is a random variable B_k with distribution function $B_k(x) = P\{B_k < x\}$.

Duration of the orientation from one user to user k is a random variable C_k with distribution function $C_k(x) = P\{C_k < x\}$. Thus C_k can be interpreted as a loss of time in preparing the service process for user of class k. The main purpose of research of polling systems is to determine the characteristics of systems development. But not always analytical formulas can be used directly, so great care is offered for numerical algorithms. In this paper, using the methodology of generalized priority systems and generalized algorithms elaborated in [3], for this characteristics numerical solutions with necessary required accuracy are obtained. Some examples and numerical results are presented.

2 The k-busy period

Definition 1. *The k-busy period is a measure of the time that expires from when a server begins to process, after an empty queue, to when the k-queue becomes empty again for the first time [3].*

Denote by Π_k^δ the length of the k-busy period, and by $\Pi_k^\delta(x) = P\{\Pi_k^\delta < x\}$, -its distribution function. Let consider that $\pi_k^\delta(s) = \int\limits_0^\infty e^{-st} d\Pi_k^\delta(x)$ is the Laplace-Stieltjes transform of distribution function of k-busy period. The proof of the presented below analytical results (Theorem 1-4) can be obtained using method of catastrophes [3].

Theorem 1. *Function $\pi_k^\delta(s)$ is determined from equation*

$$\pi_k^\delta(s) = c_k(s + \lambda_k - \lambda_k \pi_k(s))\pi_k(s), \tag{2.1}$$

where

$$\pi_k(s) = \beta_k(s + \lambda_k - \lambda_k \pi_k(s)), \tag{2.2}$$

and with $c_k(s)$ and $\beta_k(s)$ denoting the Laplace-Stieltjes transforms of distribution functions $C_k(x)$ and $B_k(x)$,

$$c_k(s) = \int\limits_0^\infty e^{-sx} dC_k(x), \;\; \beta_k(s) = \int\limits_0^\infty e^{-sx} dB_k(x).$$

Remark 2.1 *If $\lambda_k\beta_{k1} < 1$, $\lambda_k c_{k1} < 1$, then first moment of k-busy period is determined from:*

$$\pi_{k1}^{\delta} = \frac{\beta_{k1}}{1-\lambda_k\beta_{k1}} + \frac{c_{k1}}{1-\lambda_k c_{k1}}, \tag{2.3}$$

where β_{k1} and c_{k1} are the first moments of $C_k(x)$ and $B_k(x)$.

Remark 2.2 *If we consider that $C_k = 0$ and $k = 1$, then from formula (2.1) it follows that $\pi_k^{\delta}(s) = \pi_k(s)$ and $\pi_1^{\delta} = \pi_1(s) = = \beta_1(s+\lambda-\lambda\pi_1(s))$, respectively.*

Thus, expression (2.1) can be viewed as an analog of Kendall equation obtained for classical $M|G|1$ system [4].

3 Probability of states

Probability of states have a decisive role in determining important parameters such as, for example, blocking probabilities used in network technologies equipped with QoS (Quality of Service) and CoS (Class of Service). At arbitrary time $x \in (0,\infty)$ system can be in one and only one of the following states: or in serving state, or in the state of changing the states, or is free. If the polling system is in serving state, question arises: which user's message is served in that time? If the system is in the state of changing, to which class of messages the exchange occurs?

Denote $P_{B_k}(x), P_{C_k}(x)$ and $P_0(x)$ the probabilities that at instant x the system is busy by service of k-messages, switching to k-messages and system is free, respectively.

Theorem 2. *The Laplace-Stieltjes transforms of $P_{B_k}(x)$, $P_{C_k}(x)$ and $P_0(x)$ are determined from*

$$p_{B_k}(s) = \frac{\lambda_k[1-\pi_k(s)]}{s[s+\lambda_k-\lambda_k\pi_k^{\delta}(s)]}, \tag{3.1}$$

$$p_{C_k}(s) = \frac{\lambda_k[1-c_k(s)]}{s[s+\lambda_k-\lambda_k\pi_k^{\delta}(s)]}, \tag{3.2}$$

$$p_0(s) = \frac{1}{s} - [p_{B_k}(s) + p_{C_k}(s)]), \tag{3.3}$$

where $\pi_k^\delta(s)$ *and* $\pi_k(s)$ *is determined from (2.1), (2.2).*

Denote $P_{B_k} = \lim_{x\to\infty} P_{B_k}(x)$, $P_{C_k} = \lim_{x\to\infty} P_{C_k}(x)$, $P_0 = \lim_{x\to\infty} P_0(x)$.

Remark 3.1 *If* $\lambda_k\beta_{k1} < 1$, $\lambda_k c_{k1} < 1$, *then*

$$P_{B_k} = \lim_{s\downarrow 0} s p_{B_k}(s), \tag{3.4}$$

$$P_{C_k} = \lim_{s\downarrow 0} s p_{C_k}(s), \tag{3.5}$$

$$P_0 = \lim_{s\downarrow 0} s p_0(s), \tag{3.6}$$

and

$$P_{B_k} = \frac{\lambda_k \pi_{k1}}{1 + \lambda_k \pi_{k1}^\delta}, \tag{3.7}$$

$$P_{C_k} = \frac{\lambda_k c_{k1}}{1 + \lambda_k \pi_{k1}^\delta}, \tag{3.8}$$

$$P_0 = 1 - \frac{\lambda_k (c_{k1} + \pi_{k1})}{1 + \lambda_k \pi_{k1}^\delta}. \tag{3.9}$$

4 The Pollaczek-Khintchin virtual analog

Let $P_m(x)$ be the probability that at the instant x there are m messages in the k-queue. Denote

$$P_k(z, x) = \sum_{m=1}^{\infty} P_m(x) z^m, 0 \le z \le 1,$$

the generating function of queue length distribution and

$$p_k(z, s) = \int_0^\infty e^{-sx} P_k(z, x) dx,$$

the Laplace transform of function $P_k(z, s)$.

Theorem 3.

$$p_k(z,s) = \frac{1+\lambda_k \pi_k^{\delta}(z,s)}{s+\lambda_k-\lambda_k z}, \tag{4.1}$$

$$\pi_k^{\delta}(z,s) = \frac{1-c_k(s+\lambda_k-\lambda_k z)}{s+\lambda_k-\lambda_k z} + \frac{\beta_k(z,s)}{z-\beta_k(s+\lambda_k-\lambda_k z)}\times$$

$$\times[zc_k(s+\lambda_k-\lambda_k z)-\pi_k^{\delta}(s)], \tag{4.2}$$

$$\beta_k(z,s) = \frac{1-\beta_k(s+\lambda_k-\lambda_k z)}{s+\lambda_k-\lambda_k z}. \tag{4.3}$$

Remark 4.1 *In the next section it is shown that from the Theorem 3 it follows Theorem 4, where formula (5.2) results from. Thus, the result from Theorem 3 can be viewed as a virtual analogue of the Pollaczek-Khintchin equation.*

5 The Pollaczek-Khintchin steady state analog

Theorem 4. *If* $\lambda_k\beta_{k1} < 1$, $\lambda_k c_{k1} < 1$, *then*

$$P_k(z) = \lim_{s\downarrow 0} sp_k(z,s),$$

and

$$P_k(z) = \frac{1+\lambda_k\pi_k^{\delta}(z,0)}{1+\lambda_k\pi_{k1}^{\delta}}. \tag{5.1}$$

Remark 5.1 *If* $C_k = 0$ *and* $k = 1$, *then*

$$P_{k1}(z) = P_k(z) = \frac{\beta(\lambda-\lambda z)(z-1)(1-\lambda\beta_1)}{z-\beta(\lambda-\lambda z)}, \tag{5.2}$$

where $\beta_1(\cdot) = \beta(\cdot)$ *and* $\beta_{11} = \beta_1$.

Formula (5.2) is referred to in most text-books on queueing analysis and it is known as the Pollaczek-Khintchin transform equation (Pollaczek (1961); Khintchin (1963)).

6 The first moment of virtual distribution

Let $L_k(x)$ be the expectation of non stationary (virtual) distribution of k-queue length

$$l_k(s) = \int_0^\infty e^{-sx} L_k(x)dx.$$

It is easy to see that $-l_k'(s)|_{s=0} = L_k(x)$. Thus, after deriving by s the expression from Theorem 3, changing the sign and placing $s = 0$, it is obtained the analytical expression for $l_k(s)$.

Remark 6.1.

$$l_k(s) = \frac{\lambda_k}{s + \lambda_k - \lambda_k \pi_k^\delta(s)} \times$$

$$\times \{ \frac{c_{k1}\lambda_k s + \lambda_k(1 - c_k(s))}{s^2} + \frac{[1 - \beta_k(s)(s - \lambda_k) + \lambda_k s \beta_{k1}](c_k(s) - \pi_k^\delta(s))}{s^2(1 - \beta_k(s))} -$$

$$- \frac{(1 + \lambda_k \beta_{k1})(c_k(s) - \pi_k^\delta(s))}{s(1 - \beta_k(s))^2} \}. \tag{6.1}$$

7 Numerical algorithms

The analytical results formulated above, although they are of interest from fundamental theoretical point of view, are quite complicated for numerical modelling. Indeed, for example, $\pi_k^\delta(s)$ given by the Theorem 1 occurs in most of the following expressions (Theorem 2, 3, etc.). But for determining this function it is necessary to solve the functional equation (2.2), which does not have the exact analytical solution, but which effectively can be solved numerically [3, 5]. As an example, we will present a numerical algorithm of successive approximations.

Algorithm 1.

Input: $\{\lambda_k\}_{k=1}^r; \{b_k\}_{k=1}^r; \{c_k\}_{k=1}^r; s; r; \varepsilon > 0.$
Output: $k; \{\pi_k(s)\}_{k=1}^r; \{\pi_k^{\delta}(s)\}_{k=1}^r.$
Descriptions:
1. Laplace-Stieltjes transforms of exponential distribution functions $B_k(x)$ and $C_k(x)$, are determined:
$\beta_k(s) = \frac{b_k}{s+b_k}$; $\overline{c}_k = \frac{c_k}{s+c_k}$.

2. Distribution function for k-busy period is determined, using Theorem 1, for $k = 1, ..., r$. For $n = 0$,
$\pi_k^{(0)}(s) = 0$, ${\pi_k}^{(n)}(s) = \beta_k(s + \lambda_k - \lambda_k {\pi_k}^{(n-1)}(s))$,
$\pi_k^{\delta(n)}(s) = \overline{c}_k(s + \lambda_k - \lambda_k {\pi_k}^{(n)}(s)){\pi_k}^{(n)}(s)$.

Stop condition: $|\pi_k^{(n)}(s) - \pi_k^{(n-1)(s)}| < \varepsilon$.

We will further mention that the presented examples from the Section 8 are not suggested from any concrete practical examples. Their mission is to demonstrate the efficiency of the algorithms and that the elaborated algorithms are invariant to the type of the distribution function (it does not depend on concrete expression of the distribution function). We will also observe that the numerical results obtained, presented in Tables, are in full accordance and do not contradict the analytical results.

8 Numerical examples

In this section, numerical results are presented. The results of each example are obtained from the analogous algorithm, which is presented above.

- **k-busy period**
  Example 8.1 The type of distribution function taken by $B_k(x)$ and $C_k(x)$ is the *Exponential distribution*, so

$$B_k(x) = 1 - e^{b_k x}, x > 0 \text{ and } C_k(x) = 1 - e^{c_k x}, x > 0,$$

with the following parameters:

λ_k ={2, 3, 5, 6, 3, 8, 5, 4, 7, 3, 4, 2, 9, 7, 6, 4, 6, 3, 5, 2},

b_k ={7, 5, 9, 4, 6, 2, 5, 4, 8, 6, 5, 3, 4, 5, 7, 6, 2, 4, 8, 6},

c_k ={7, 5, 9, 4, 6, 2, 5, 4, 8, 6, 5, 3, 4, 5, 7, 6, 2, 4, 8, 6},

s =0.2.

The results of the program are presented in Table 1.

Table 1. Numerical results for LST of distribution function for k-busy period $\pi_k^\delta(s)$

k	$\pi_k(s)$	$\pi_k^\delta(s)$	k	$\pi_k(s)$	$\pi_k^\delta(s)$
1	0.499831	0.181751	11	0.369369	0.096169
2	0.389415	0.151739	12	0.287740	0.096169
3	0.515323	0.098878	13	0.236751	0.075282
4	0.278846	0.121721	14	0.316054	0.085492
5	0.438095	0.149959	15	0.424283	0.180499
6	0.133333	0.042254	16	0.418269	0.110781
7	0.350425	0.122974	17	0.152174	0.047297
8	0.313589	0.058366	18	0.333156	0.155468
9	0.445795	0.199467	19	0.480989	0.197734
10	0.438095	0.172673	20	0.458472	0.165468

Example 8.2 The type of distribution function taken by $B_k(x)$ is the *Erlang distribution* and by $C_k(x)$ is the *Exponential distribution*, so

$$B_k(x) = \begin{cases} 0, & x < 0 \\ \int\limits_0^x \lambda_k \frac{(\lambda_k u)^{k-1}}{(k-1)!} e^{-\lambda_k u} du, & x \geq 0 \end{cases}$$

and

$$C_k(x) = 1 - e^{c_k x}, x > 0,$$

with the following parameters:

λ_k ={3, 5, 7, 6, 8, 4, 5, 7, 3, 6, 8, 5, 9, 7, 6, 2, 4, 1, 1, 1},

b_k ={4, 4, 6, 7, 4, 8, 8, 3, 2, 5, 5, 7, 6, 4, 8, 6, 4, 2, 4, 6},

c_k ={6, 7, 4, 3, 7, 6, 9, 7, 4, 5, 3, 2, 5, 6, 7, 5, 4, 3, 2, 8},

p_k ={7, 5, 4, 3, 6, 5, 2, 5, 5, 7, 6, 5, 9, 8, 6, 5, 4, 6, 7, 6},

s =0.2.

The results of the program are presented in Table 2.

Table 2. Numerical results for LST of distribution function for k-busy period $\pi_k^\delta(s)$

k	$\pi_k(s)$	$\pi_k^\delta(s)$	k	$\pi_k(s)$	$\pi_k^\delta(s)$
1	0.000155	5.828645e-05	11	0.000244	4.069452e-05
2	0.000977	0.000360	12	0.006788	0.000972
3	0.008100	0.001806	13	8.347513e-06	1.98751e-06
4	0.044573	0.008499	14	5.947027e-06	1.784112e-06
5	8.706395e-05	2.770304e-05	15	0.424283	0.001071
6	0.013420	0.004751	16	0.010310	0.003688
7	0.174848	0.078190	17	0.0050	0.001350
8	0.000171	5.705125e-05	18	6.4e-05	1.745465e-05
9	0.000129	3.67441e-05	19	0.000457	9.145366e-05
10	0.000128	3.544855e-05	20	0.006196	0.003099

Example 8.3 The type of distribution function taken by $B_k(x)$ is the *Erlang distribution* and by $C_k(x)$ is *Normal distribution*, so

$$B_k(x) = \begin{cases} 0, & x < 0 \\ \int\limits_0^x \lambda_k \frac{(\lambda_k u)^{k-1}}{(k-1)!} e^{-\lambda_k u} du, & x \geq 0 \end{cases}$$

and

$$C_k(x) = \frac{1}{\sigma_k}\sqrt{2\pi} \int\limits_{-\infty}^{x} e^{-\frac{(u-a)^2}{2\sigma_k{}^2}} du, \ -\infty < x < \infty,$$

with the following parameters:

λ_k ={0.6, 0.4, 0.6, 0.7, 0.3, 0.4, 0.6, 0.2, 0.6, 0.4, 0.7, 0.1, 0.3, 0.5, 0.7, 0.5},

b_k ={0.2, 0.4, 0.7, 0.4, 0.6, 0.8, 0.6, 0.3, 0.4, 0.5, 0.3, 0.2, 0.5, 0.4, 0.6, 0.7},

c_k ={0.5, 0.3, 0.5, 0.7, 0.5, 0.3, 0.4, 0.8, 0.5, 0.4, 0.2, 0.3, 0.4, 0.5, 0.3, 0.5},

p_k ={2, 4, 3, 5, 4, 6, 5, 4, 3, 2, 4, 6, 5, 4, 7, 6},

σ_k ={2, 4, 3, 5, 4, 6, 5, 4, 3, 2, 4, 6, 5, 4, 7, 6}, s =0.2.

The results are presented in the Table 3.

Table 3. Numerical results for LST of distribution function for k-busy period $\pi_k^\delta(s)$

k	$\pi_k(s)$	$\pi_k^\delta(s)$	k	$\pi_k(s)$	$\pi_k^\delta(s)$
1	0.028566	0.019760	9	0.024037	0.015351
2	0.012551	0.011409	10	0.163897	0.154544
3	0.075241	0.058980	11	0.002108	0.001942
4	0.001348	0.000656	12	0.000544	0.000473
5	0.047328	0.034053	13	0.012758	0.010210
6	0.015996	0.015313	14	0.008963	0.007629
7	0.007416	0.00564	15	0.000682	0.000563
8	0.012482	0.008618	16	0.007012	0.005264

Example 8.4 The types of distribution function for $B_k(x)$ and $C_k(x)$ are given in the Table 5. We used the following notations:

- If the distribution function is *Uniform* we denote it by the letter **U**,
- If the distribution function is *Erlang* – **I**,
- If the distribution function is *Exponential* – **E**,
- If the distribution function is *Normal* – **N**.

λ_k ={0.2, 0.4, 0.7, 0.5, 0.6, 0.3, 0.4, 0.5, 0.6, 0.9},

s =0.2.

Required parameters for each distribution function are given in the Table 4.

Table 4. Parameters for distribution functions

U	N	U	I	N
$a = 4$ $b = 1$	$\sigma = 0.6$ $a = 0.2$	$a = 5$ $b = 2$	$k = 2$	$\sigma = 0.9$ $a = 0.2$
U	I	N	I	U
$a = 10$	$k = 3$	$\sigma = 0.4$ $b = 2$	$k = 4$ $a = 0.1$	$a = 14$ $b = 7$
I	U	I	U	I
$k = 1$	$a = 9$ $b = 2$	$k = 2$	$a = 3$	$k = 2$ $b = 1$

The results of the program are given in the Table 5.

Table 5. Numerical results for LST of distribution function for k-busy period $\pi_k^\delta(s)$

k	$B_k(x)$	$C_k(x)$	$\pi_k(s)$	$\pi_k^\delta(s)$
1	U	N	0.367026	0.363908
2	E	U	0.846724	0.250955
3	I	N	0.599745	0.675492
4	E	U	0.894586	0.147271
5	I	N	0.608076	0.601602
6	I	U	0.68603	0.014749
7	E	I	0.704387	0.496661
8	E	U	0.920868	0.178205
9	I	E	0.624615	0.565239
10	U	I	0.139467	0.059018

For the case when $B_k(x)$ and $C_k(x)$ are PH distribution functions, it is necessary to obtain the matrix form of Kendall and generalized Kendall equations. These analitical results are obtained in [5].

Example 8.5 The types of distribution functions taken by $B_k(x)$ and $C_k(x)$ are PH distributions with representation (α^t, T_k), (α^t, P_k), so

$$B_k(x) = 1 - \alpha_t e^{T_k x} e, x > 0,$$

$$C_k(x) = 1 - \alpha_t e^{P_k x} e, x > 0,$$

with the following parameters:

$\lambda_k = \{0.4; 0.4; 0.3; 0.6; 0.5\}$,

$\tilde{\lambda}_k = \{0.4; 0.2; 0.4; 0.6; 0.6\}$,

$\delta_k = \{0.6; 0.3; 0.4; 0.5; 0.2\}$,

$s = 0.2$.

The results of the program are presented in Table 6.

Table 6. Numerical results for LST of distribution function for k-busy period $\pi_k^\delta(s)$

k	$\pi_k(s)$	$\pi_k^\delta(s)$
1	0.026725	0.901550
2	0.066859	0.718482
3	0.012627	0.928276
4	0.035894	0.804515
5	0.022830	0.709058

Example 8.6 The types of distribution functions taken by $B_k(x)$ and $C_k(x)$ are PH distributions with representation (α^t, T_k), (α^t, P_k), so

$$B_k(x) = 1 - \alpha_t e^{T_k x} e, x > 0,$$

$$C_k(x) = 1 - \alpha_t e^{P_k x} e, x > 0,$$

with the following parameters:

$\lambda_k = \{0.5; 0.6; 0.3; 0.4; 0.5; 0.2; 0.6; 0.6; 0.2; 0.1\}$,

$\tilde{\lambda_k} = \{0.2; 0.3; 0.4; 0.2; 0.6; 0.7; 0.8; 0.4; 0.2; 0.3\}$,

$\delta_k = \{0.3; 0.4; 0.1; 0.2; 0.6; 0.8; 0.5; 0.4; 0.4; 0.8\}$,

$s = 0.5$.

The results of the program are presented in Table 7.

Table 7. Numerical results for LST of distribution function for k-busy period $\pi_k^\delta(s)$

k	$\pi_k(s)$	$\pi_k^\delta(s)$	k	$\pi_k(s)$	$\pi_k^\delta(s)$
1	0.012692	0.864672	6	0.000060	0.999554
2	0.014688	0.865907	7	0.003161	0.959972
3	0.000978	0.957140	8	0.010383	0.891422
4	0.006395	0.895401	9	0.000542	0.995270
5	0.002997	0.973018	10	0.000017	0.999915

Example 8.7 The types of distribution functions taken by $B_k(x)$ and $C_k(x)$ are PH distributions with representation (α^t, T_k), (α^t, P_k), so

$$B_k(x) = 1 - \alpha_t e^{T_k x} e, x > 0,$$

$$C_k(x) = 1 - \alpha_t e^{P_k x} e, x > 0,$$

with the following parameters:

$\lambda_k = \{0.2; 0.3; 0.1; 0.5; 0.6; 0.7; 0.4; 0.8; 0.4; 0.5; 0.3; 0.7; 0.8; 0.4;$
$0.6; 0.9; 0.3; 0.4; 0.5; 0.4\}$,

$\tilde{\lambda_k} = \{0.2; 0.3; 0.5; 0.2; 0.3; 0.7; 0.8; 0.4; 0.3; 0.5; 0.1; 0.5; 0.8; 0.4;$
$0.3; 0.6; 0.4; 0.9; 0.4; 0.2\}$,

$\delta_k = \{0.5; 0.4; 0.8; 0.4; 0.4; 0.7; 0.4; 0.3; 0.8; 0.2; 0.1; 0.5; 0.4; 0.7;$
$0.5; 0.4; 0.7; 0.9; 0.4; 0.3\}$,

$s = 0.5$.

The results of the program are presented in Table 8.

Table 8. Numerical results for LST of distribution function for k-busy period $\pi_k^\delta(s)$

k	$\pi_k(s)$	$\pi_k^\delta(s)$	k	$\pi_k(s)$	$\pi_k^\delta(s)$
1	0.002444	0.986440	11	0.012414	0.750670
2	0.005566	0.956771	12	0.029777	0.796091
3	0.000068	0.999659	13	0.021775	0.763410
4	0.030767	0.812228	14	0.009064	0.954890
5	0.034760	0.766861	15	0.034760	0.807550
6	0.018947	0.881203	16	0.043203	0.644660
7	0.003083	0.960978	17	0.003927	0.980092
8	0.051492	0.569886	18	0.002451	0.984919
9	0.012501	0.951740	19	0.016581	0.864629
10	0.012555	0.785024	20	0.017712	0.851134

- **Probability of states**

Example 8.8 $k = 4$, $\lambda_4 = 1.456$. Time of serving is Erlang with parameters $\alpha = 1$ and $k_e = 2$, $Erl(1, 2)$. Time of exchange for user k is Exponential with parameter $c = 0.268, E(0.268)$. For different values of t we have the probabilities $P_{B_4}(t)$ and $P_{C_4}(t)$. The results are presented in Table 9 and Table 10.

Table 9. Numerical results for probabilities of states for the 4th user, $\lambda_4 = 1.456$, in terms of LST and different values of t

t	$P_{B_4}(t)$	$P_{C_4}(t)$
0.01	0.833	0.0193
0.2	0.07599	0.00125
0.5	0.06629	0.00116
1.5	0.04472	0.00049
2.9	0.02874	0.00019
5.8	0.01448	0.00005
10.4	0.00680	0.00001

Table 10. Numerical results for probabilities of states for the 4th user, $\lambda_4 = 1.456$, in terms of LST

t	$P_{B_4}(t)$	$P_{C_4}(t)$
1	0.05389	0.00074
2	0.037774	0.00034
3	0.02794	0.00018
4	0.2153	0.0010
5	0.01711	0.00006
6	0.01396	0.00009
7	0.01155	0.00003
8	0.00974	0.00002
9	0.00832	0.00001
10	0.00719	0.00001

Example 8.9 We have the same parameters like in Example 8.8 with the exception of flow requirements parameter for user k: $\lambda_k = 2.987465$. The results are presented in Table 11 and Table 12.

Table 11. Numerical results for probabilities of states for the 4th user, $\lambda_4 = 2.987465$, in terms of LST and different values of t

t	$P_{B_4}(t)$	$P_{C_4}(t)$
0.01	0.06466	0.03934
0.2	0.06200	0.03217
0.5	0.05781	0.02413
1.5	0.04553	0.01101
2.9	0.03336	0.00457
5.8	0.01957	0.00133
10.4	0.01037	0.00033

Table 12. Numerical results for probabilities of states for the 4th user, $\lambda_4 = 2.987465$, in terms of LST

t	$P_{B_4}(t)$	$P_{C_4}(t)$
1	0.05128	0.01587
2	0.04057	0.00795
3	0.03268	0.00451
4	0.02682	0.00277
5	0.02238	0.00181
6	0.01895	0.00124
7	0.01625	0.00088
8	0.014909	0.00064
9	0.01233	0.00048
10	0.01088	0.00036

Remark 8.1 *Theorem 2 gives us the possibility for modeling the probability of states for an arbitrary* $k, 1 \leq k \leq r$.

- **Queue length**

Example 8.10 Let $n = 9$, where n is the number of users and parameters λ_k, b_k and c_k are presented in the form of vectors:

$\lambda_k = \{0.2, 0.4, 0.3, 0.1, 0.5, 0.6, 0.4, 0.1, 0.2\}$,

$b_k = \{0.1, 0.4, 0.5, 0.1, 0.2, 0.4, 0.1, 0.5, 0.3\}$,

$c_k = \{0.4, 0.6, 0.4, 0.2, 0.1, 0.2, 0.3, 0.4, 0.1\}, s = 0.2.$

The results are presented in Table 13.

Table 13. Numerical results for queue length distribution $l_k(s)$, in terms of LST

k	$\pi_k(s)$	$\pi_k^{\delta}(s)$	$l_k(s)$
1	0.21904	0.11586	5.52940
2	0.49853	0.29894	3.86539
3	0.61134	0.34125	1.63709
4	0.26785	0.11320	2.87008
5	0.25865	0.03856	3.01735
6	0.42105	0.11267	1.91095
7	0.15679	0.05618	5.32859
8	0.68287	0.43239	1.04412
9	0.49806	0.12439	2.17897

References

[1] V. M. Vishnevsky and O. V. Semenova, *Polling Systems: The theory and applications in the broadband wireless networks.* Moscow, Russia: Texnocfera, 2007, 312 p. (in Russian)

[2] V. V. Rycov and Gh. K. Mishkoy, "A new approach for analysis of polling systems," in *Proc. of the Int. Conf. on Control Problems*, Moscow, 2009, pp. 1749–1758.

[3] Gh. K. Mishkoy, *Generalized Priority Systems.* Chisinau, Republic of Moldova: Stiinta, 2009, 200 p. (in Russian)

[4] D. G. Kendall, "Some problems in the theory of queues," *J. Roy. Statist. Soc. (B).*, vol. 13, no. 2, pp. 151–185, 1951.

[5] Gh. Mishkoy, Udo R. Krieger and D. Bejenari, "Matrix algorithm for Polling models with PH distribution," *Buletinul Academiei de Stiinte a Republicii Moldova. Matematica*, vol. 68, no. 1, pp. 70–80, 2012.

Gh. Mishkoy, D. Bejenari,
L. Mitev, I. R. Ticu

Received June 01, 2016

Gh. Mishkoy
Academy of Science of Moldova, Free International University of Moldova
Chisinau, Republic of Moldova
E–mail: gmiscoi@ulim.md

D. Bejenari
Free International University of Moldova
Chisinau, Republic of Moldova
E–mail: artemis85@mail.ru

L. Mitev
Institute of Mathematics and Computer Science of Academy of Science of Moldova,
Free International University of Moldova
Chisinau, Republic of Moldova
E–mail: liliausate@yahoo.com

I. R. Ticu
Constanta Maritime University
Constanta, Romania
E–mail: ionela.ticu@yahoo.com

Computer Science Journal of Moldova, vol.**24**, no.2(71), 2016

Digital Health Data: A Comprehensive Review of Privacy and Security Risks and Some Recommendations *

Shahidul Islam Khan, Abu Sayed Md. Latiful Hoque

Abstract

In todays world, health data are being produced in ever-increasing amounts due to extensive use of medical devices generating data in digital form. These data are stored in diverse formats at different health information systems. Medical practitioners and researchers can be benefited significantly if these massive heterogeneous data could be integrated and made accessible through a common platform. On the other hand, digital health data containing protected health information (PHI) are the main target of the cybercriminals. In this paper, we have provided a state of the art review of the security threats in the integrated healthcare information systems. According to our analysis, healthcare data servers are leading target of the hackers because of monetary value. At present, attacks on healthcare organizations' data are 1.25 times higher compared to five years ago. We have provided some important recommendations to minimize the risk of attacks and to reduce the chance of compromising patients' privacy after any successful attack.

Keywords: Health Data, Privacy, Security, Data Breach, PHI

1 Introduction

Health data refers to pieces of information collected to use in the diagnosis of a health condition. Health Information is collected about

* This research is supported by the ICT Division, Ministry of Posts, Telecommunication and Information Technology, Government of the People's Republic of Bangladesh.

a patient, his/ her family, often during creating of a nursing history for the patient. A health record may include multiple types of health data such as various notes entered by health care professionals over time, recording observations and administration of drugs, test results, x-rays, reports, etc. Digital health data are health data generated by medical devices in digital form e.g., fasting plasma glucose test (FGT) result, or other patient health related information e.g., height, weight, blood group etc stored in digital form at computers, laptops, or in database of health information systems [1]–[3].

At present, enormous quantity of digital health data are generated daily by healthcare providers. Medical records of patients are increasingly digital, in the form of Electronic Health Record (EHR). These EHRs are more useful than paper records for better healthcare and medical research because electronic data can be stored easily and manipulated by software. These precious data are stored in various health information systems (HIS) in hospitals, research centers and diagnostic laboratories. Many of these data fall in the category of protected health information.

Protected health information (PHI) is defined as personally identifiable health information collected from an individual, and covered under federal or international data breach disclosure laws [4]. PHI of an Individual is information which relates to:

a. the individuals past, present, or future physical or mental health or condition,

b. the provision of health care to the individual,

c. the past, present, or future payment for the provision of health care to the individual, and that identifies the individual or for which there is a reasonable basis to believe that the information could be used to identify the individual.

PHI includes many common identifiers such as name, date of birth, address, National ID / Social Security Number, telephone and fax numbers, E-mail addresses etc. when they can be associated with the health information listed above [5].Laboratory reports, medical records, and

hospital bills are examples of PHI because each document contains a patient's name and/or other identifying information associated with the health data content.

Security of a HIS deals with protecting medical data from intruders, malwares, and frauds. It retains confidentiality and integrity of healthcare data. Privacy concerns exist wherever personally identifiable information or other sensitive information is collected and stored in any form. A major challenge in health data privacy is to share data among medical practitioners while protecting personally identifiable information. Information privacy may be applied in numerous ways, including encryption, authentication and data masking – each attempting to ensure that information is available only to authorized persons [6], [7].

Nowadays, hacking PHI by cybercriminals is observed as a growing trend. Hackers goal is to take advantage of personal information of the patients. Average sell value of a complete medical record varies from $10 to $1,000 in black market. Although privacy of a patient can be compromised with paper based medical records, it alarmingly increased along with digitized record keeping by the healthcare providers [8], [9].

It is obvious that developing a national health data warehouse (NHDW), where integrated data from all the diverse HIS will be made available for better health delivery and medical research, is very much essential for every country [10]–[16]. However NHDW raises high risk to data security and privacy of individuals. Before integration to NHDW, sensitive and private data of patients reside to a single organization such as a hospital or a diagnostic center. Only that particular organization is responsible by law to protect the data privately. Now the situation is far different in the case of national warehouse. So proper measures have to be taken to safeguard privacy of patients in the NHDW.

In this paper we have presented a comprehensive review of security and privacy risks of digital health data and integrated health information systems. We have exposed the statistics of high rise of security threads in healthcare data servers. In addition, we have provided some general recommendations to reduce risks of PHI breaches and some specific recommendations for developing national scale integrated health

information systems.

2 Data Breaches of Health Information Systems

A health data breach or leakage is defined as an event that involves the loss or exposure of personal health records. Personal health records are data containing privileged health related information about an individual that cannot be readily obtained through other public means, which information is only known by an individual or by an organization under the terms of a confidentiality agreement [17]. For example, leakage of a health insurer's record of the policyholder with doctor and payment information will be treated as a health data breach. According to the research by IBM and Ponemon Institute in 2015 where 350 companies in 11 countries were interviewed extensively, more than 18 thousand records were breached on an average in each breached incident [18]. This is presented in Fig. 1.

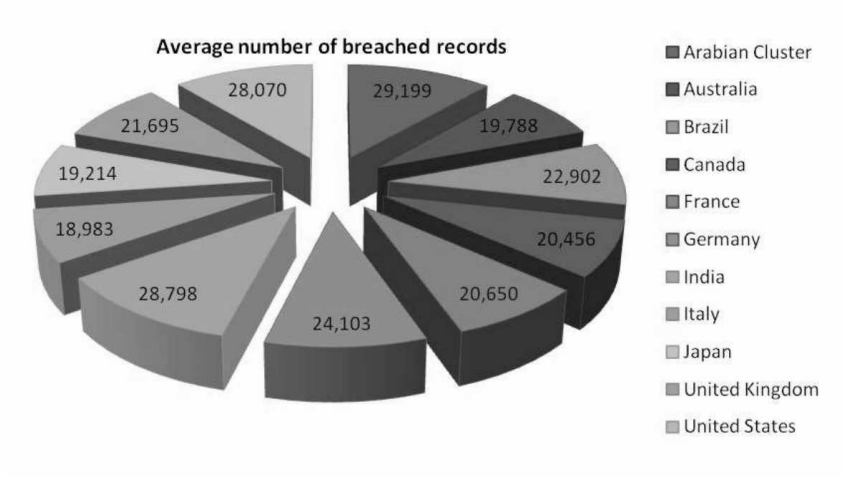

Figure 1. Average number of breached records in a data breach incident

The costs of a data breach can vary according to the cause and the protections in place at the time of the breach. Direct costs refer to the direct expense spent to carry out a given activity such as hiring forensic experts and law firm or offering identity protection services to the victims. Indirect costs include the time, effort and other organizational resources spent during the data breach resolution. Indirect costs also include the loss of goodwill and customer churn. In 2015, the average cost of data breach per lost or stolen record was 154USD but in case of a breach of healthcare organization, the average cost was 363USD [18]. This is shown in Fig. 2.

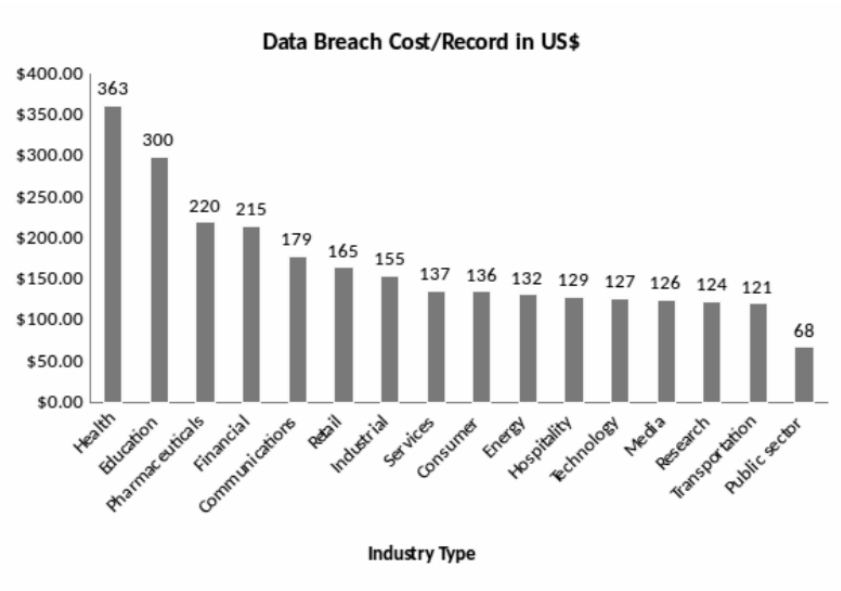

Figure 2. Cost of each breached record in different sector. The cost is maximum for the healthcare industry.

2.1 Health data breaches

According to 2015 Fifth Annual Benchmark Study on Privacy and Security of Healthcare Data which covered 90 healthcare organizations in USA, more than 90% of healthcare service providers had a data breach, and 40% had more than five data breaches over the past two years [19]. The following chart of Fig. 3 shows the total numbers of health data breaches in USA in last five years till February 26, 2016. We have calculated the data from [20].

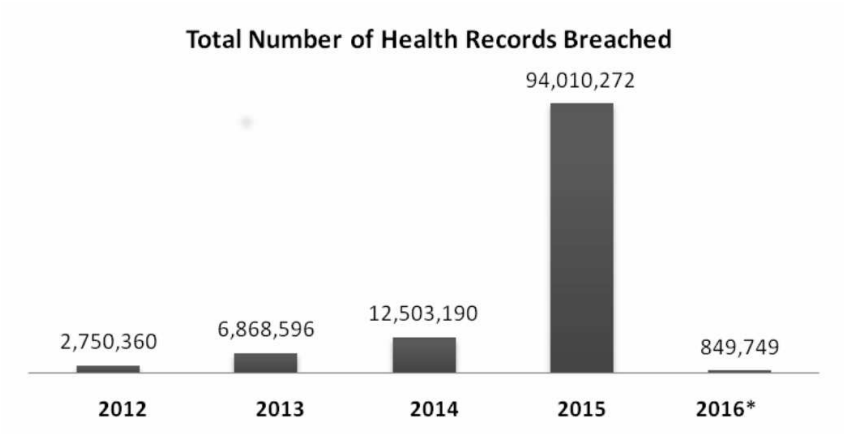

Figure 3. Total number of health records breached in USA

According to the report [19], for the first time, criminal attacks are the number one cause of healthcare data breaches. Criminal attacks on healthcare organizations are 1.25 times higher compared to five years ago. The main causes of data breach in healthcare sectors are illustrated in Fig. 4.

Some recent attacks on health information centers are listed below:

- Hackers have shut down the internal computer system at a Hollywood Presbyterian Medical Center for more than a week for a payoff of 9,000 bitcoins, or almost USD 3.7 million [21]. It is due to a malicious software called ransomware that encrypts sensitive data until it can only be decrypted with a code.

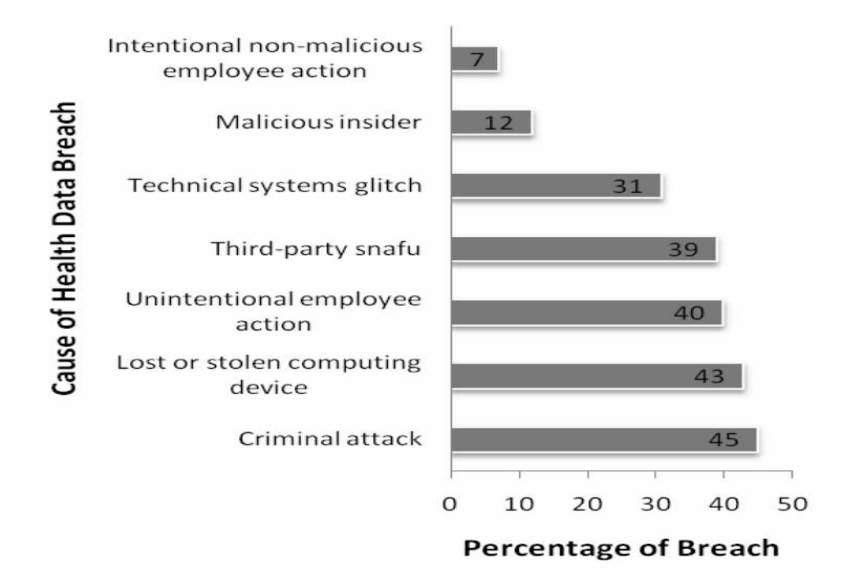

Figure 4. Main causes of data breach in the healthcare industry

- In February 2016, Jackson Health System discovered that a hospital employee have stolen confidential PHI of patients including names, birthdates, social security numbers and home addresses around 24,000 patient records over the last five years [22].

- The Washington State HCA reported, in February 2016, that an employee error resulted in a healthcare data breach compromising 91,000 Medicaid patient files. The information affected includes clients social security numbers, dates of birth, Apple Health client ID numbers and private health information [23].

- Six hard drives containing personal and health information on clients of health insurance company Centene Corp were lost which contained Social Security numbers, birthdates, health data, names, addresses, and insurance identification numbers for 950,000 patients who received laboratory services between 2009-2015 [24].

- Premera Blue Cross was targeted with a sophisticated cyber attack after hackers gained access to the financial and medical information of 11 million members in January 2015. Hackers swiped Social Security numbers, financial information, medical claims data, addresses, email addresses, names and dates of birth [25].

- Health insurer Anthem Inc. has suffered a massive data breach on March 3, 2015 after hackers gained access to a corporate database reportedly containing personal information on around 80 million of the health insurer's current and former USA customers and employees [26].

- In last ten years at least 18 health breaches reported in Europe affected minimum 9,337,197 individual records [17]. The health records include details on the patients conditions, names, home addresses and dates of birth. The health networks and servers containing integrated health records are in high risk of cyber attacks all over the world.

2.2 Data breaches of healthcare servers

From 2014, hackings on healthcare servers increased terrifyingly. The attackers motivation is to get huge PHI in a single successful hack. Table 1 presents last 12 big criminal attacks on integrated health records in USA within last 12 months. We have summarized these data from [20].

We have analyzed the data provided by U.S. Department of Health and Human Services and found that hackers are increasingly targeted healthcare servers which is very alarming to national level health information system development. Table 2 and Fig. 5 illustrate the fact clearly.

2.3 Other impacts of health data breaches

There are other impacts of health data breaches. They are discussed below:

Table 1. Latest 12 big breaches in USA on Health Data Servers

Sl.	Name of Healthcare Org.	Affected Individuals	Breach Date	Type of Breach
1	Alliance Health Networks, LLC	42372	2/15/2016	Hacking/IT Incident
2	OH Muhlenberg, LLC	84681	11/13/2015	Hacking/IT Incident
3	Excellus Health Plan, Inc.	10000000	9/9/2015	Hacking/IT Incident
4	Medical Informatics Engineering	3900000	7/23/2015	Hacking/IT Incident
5	University of California, Los Angeles Health	4500000	7/17/2015	Hacking/IT Incident
6	CareFirst BlueCross BlueShield	1100000	5/20/2015	Hacking/IT Incident
7	Freelancers Insurance Company	43068	3/24/2015	Hacking/IT Incident
8	ATnT Group Health Plan	50000	3/23/2015	Hacking/IT Incident
9	Premera Blue Cross	11000000	3/17/2015	Hacking/IT Incident
10	Anthem, Inc. Affiliated Covered Entity	78800000	3/13/2015	Hacking/IT Incident
11	Virginia (VADMAS)	697586	3/12/2015	Hacking/IT Incident
12	Georgia Department of Community Health	912906	3/2/2015	Hacking/IT Incident

Table 2. Statistics of Healthcare server attack compared to total healthcare breach

Reporting Year	Total Health Data Breach affecting 500 or more individuals	Healthcare Server Attach
January 1, 2011 to December 31, 2011	194	27
January 1, 2012 to December 31, 2012	202	25
January 1, 2013 to December 31, 2013	263	35
January 1, 2014 to December 31, 2014	290	55
January 1, 2015 to December 31, 2015	265	50

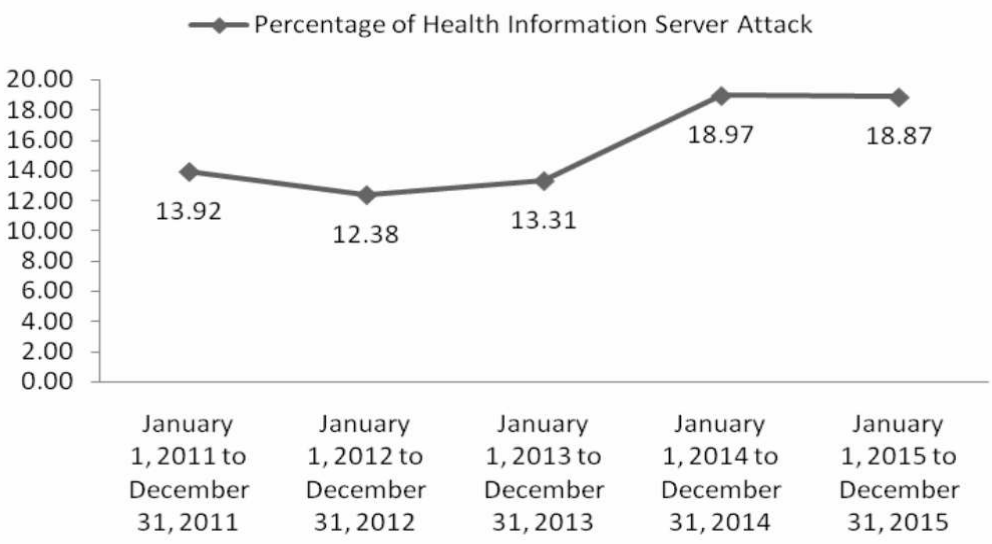

Figure 5. Criminal attack on Healthcare data servers are increasing high.

a. Breaches of PHI drastically effect on the goodwill of a healthcare organization. In a research report it is shown that, people are withholding their health information from healthcare providers because they are concerned that there could be a confidentiality breach of their records [27]. An unwillingness to fully disclose information could delay a diagnosis of a communicable disease. This is not only a potential issue for the treatment of a specific patient; there are potential public health implications.

b. Penalty of healthcare providers are imposed in two ways. They have to pay ransom to the hackers to get their breached data back or to restore their hacked system [21] and they also pay the government privacy penalty for failing to safeguard patient information [28].

3 Analysis of the risks related to Health Information Systems

If we analyze the increase trend of healthcare data breach around the globe, it becomes quite clear that the main reason of the breaches is the sell value of complete health records. What makes medical data so unique is that it often contains most of the information hackers are looking for such as credit card information, and Social Security and bank account numbers giving them a one-stop stealing strategy. Fraudsters use this data to create fake IDs to buy medical equipment or drugs that can be resold, or they combine a patient number with a false provider number and file made-up claims with insurers. Sometimes the cyber criminals use this data to blackmail a patient with good social status. For example, F1 racing legend Michael Schumachers and pop legend Michael Jacksons medical records were hacked.

If we look at Table 1, we can see that, all big breaches in healthcare servers are cause of hacking or IT incident though there are other causes available in the U.S. Govt. reporting form i.e., Theft, Unauthorized Access/Disclosure, Lossor unknown cause. So the owner of the healthcare servers should pay high attention to develop a secure

framework to protect their health information servers from hacking or improper IT involvements.

Another important thing to notice is that, a healthcare company is looser in many ways after a successful breach. It has to pay money to both the hackers and the government. This situation will eventually increase healthcare cost and decrease better healthcare delivery. Policymaker should think about this.

If the stored health data are de-identified in every place from health information system software to backups and also in health data warehouses, then the risk of data breach can be significantly reduced. Because there is almost no sell value of de-identified health records. Another positive thing of de-identification is if a data breach occurs, privacy of individual patient will not be affected.

4 Some general recommendations to reduce the chance of health data breaches

a. At the very least, healthcare companies should back up all their important health data regularly so that, in emergency situations, hard drives can be cleaned and restored to their previous states. PHIs in database backups must also be encrypted.

b. Internal HIS software should be screened for loopholes that could be way in of hackers. All third party software should be updated with latest patch and service packs. No free software from unknown or un-trusted source should ever be downloaded or installed.

c. Doctors and nurses should be more careful when handling PHI of patients. They should encrypt these records in their own laptops and pen drives. After working in the workstations, they must always sign out from their accounts when they have finished inputting patient information or viewed patients reports.

d. Health-care consumers should be smarter. The more the patients will query healthcare providers about how they are securing PHI,

the more attention the providers will pay to enhance security and privacy of patients PHI.

e. It is more effective to integrate privacy and security into health apps, devices, and services from the start. For any piece of information collection and storage, the following should be considered:

   i. Minimize the amount of personal information collected
   ii. Decide how long the information needs to be stored
   iii. Encrypt information when possible
   iv. Delete the information earliest

f. Rather than spending a lot of money after breaches, the healthcare organizations should increase their budget for HIS security. Prevention is better than cure- this proverb should always be remembered.

g. Medical practitioners need to be more cautious of email attachments and shouldnt include health information in e-mail unless encryption is used. If encryption is not available, confidentiality statement needs to be included like below at the top of the e-mail:

> **Notice: Privacy & Confidentiality of Information**
> *This communication may contain non-public, confidential, or legally privileged information intended for the sole use of the designated recipients. If you are not the intended recipient, or have received this communication in error, please notify the sender immediately by reply email at xxx@xxx.xx or by telephone at +xxx-xxxxxxx, and delete all copies of this communication, including attachments, without reading them or saving them to disk. If you are the intended recipient, you must secure the contents in accordance with all applicable state or federal requirements related to the privacy and confidentiality of information, including the HIPAA/ EU Data Protection Directive Privacy guidelines.*

5 Specific Recommendations for Deployment of National Health Data Warehouse

No information system can be assumed to be completely protected from all kind of criminal and cyber attacks. Security can be more vulnerable in the case of large scale, national level health information systems where Internet communication has to be maintained for the sake of easy data collection from far-most parts of the country. So integrated health information systems should be designed in such a way that:

- There is enough data to maintain record linkage so that doctors, researchers can get useful insight from the system.
- If data breach occurs, individual patients privacy will be safeguarded.

Record linkage is the process of identifying record pairs from different information systems which belong to the same real world entity. Given two repositories of records, the record-linkage process consists of determining all pairs that are similar to each other. Record linkage is essential when joining datasets based on entities that may or may not share a common identifier such as national id or social security number [29], [30]. For discovering effective knowledge such as correlations among diseases from medical dataset it is very essential to maintain record linkage. On the other hand, identifiable health data have high risk to patient privacy and make the health information systems security vulnerable to hackers [31], [32] For development of national level health data warehouse our recommendations from security and privacy point of view are:

1. No Medical record can be stored in any level, from diagnostic centers to National Health Data Warehouse, with personal identifiable attributes of the patients.
2. To facilitate knowledge discovery process of the Healthcare researchers, sufficient record-linkage data have to be kept in medical

records by replacing personal identifiable attributes with unique code using suitable computer cryptographic technique.

3. A data-protection strategy has to be implemented that will cover data everywhere it is stored, and at every stage, from creation and processing, to storage, backup and transmission.

4. Proper security measures have to be taken and tested before connecting the national health data warehouse with Internet.

5. Proper security measures have to be taken and tested before deploying the national health data warehouse in the public cloud.

We propose the following flow chart that will significantly reduce cyber attack in the national health data warehouse and also retain the privacy of the patients after any data breach incident shown in Fig. 6.

6 Conclusions

Widespread use of digital health data could bring positive changes to the healthcare system in a various ways, as these data are the foundational piece to softwares and technologies that could advance health care delivery radically. Having every patient's data stored digitally, in a national platform creating an easy transfer and comparison of data among providers, insurers, and researchers, will allow recognition of interesting medical patterns, development of personalized and predictive medicine, reductions in medical errors, better disease management, predicting and preventing disease outbreaks, elimination of insurance fraud, identification of low cost treatments and many more. However integration of protected health information has high risk to patients' privacy and makes such systems vulnerable to hackers. In this paper, we have provided a state of the art review of security and privacy risks of integrated healthcare information system. We have analyzed current security and privacy threats and provided some recommendations to reduce health data breaches. We have also provided some guidelines for developing national scale integrated health information systems.

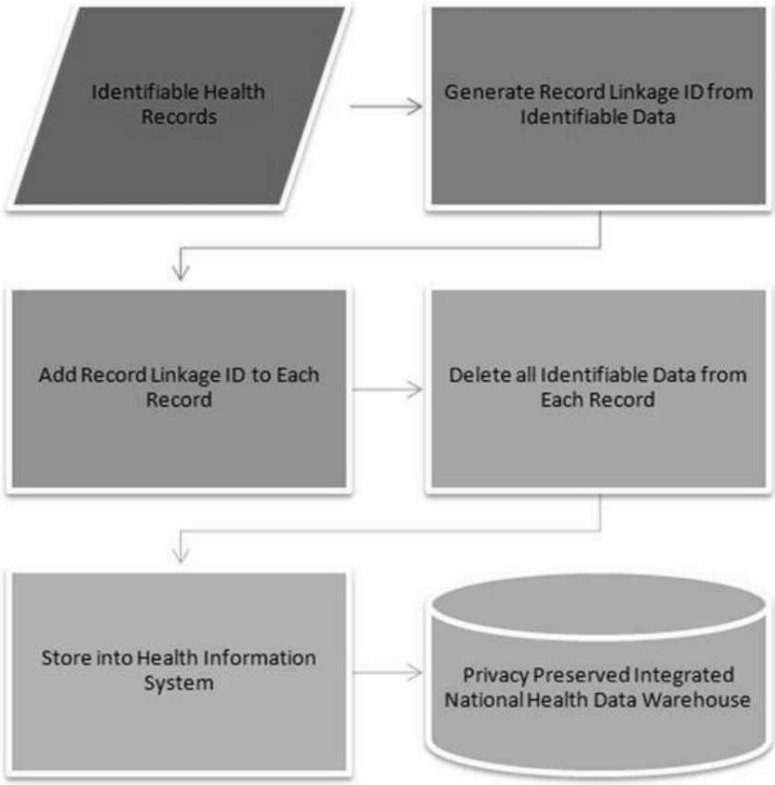

Figure 6. Flow chart of security and privacy management of National health data Warehouse

References

[1] C. K. Reddy and C. C. Aggarwal, *Healthcare data analysis.* CRC Press, 2015.

[2] Y. Zhang and C. Poon, "Editorial note on bio, medical and health informatics," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, pp. 543–545, 2010.

[3] M. L. Braunstein, *Practitioners Guide to Health Informatics.*

Springer, 2015.

[4] (2016, Feb.). [Online]. Available: http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

[5] (2016, Feb.) Protected health information: What does phi include? [Online]. Available: https://www.hipaa.com/hipaa-protected-health-information-what-does-phi-include

[6] F. T. Harold and K. Micki, *Information Security Management Handbook*, 6th ed. CRC Press, 2015, vol. 2.

[7] B. P. Robichau, *Healthcare Information Privacy and Security: Regulatory Compliance and Data Security in the Age of Electronic Health Records*, 1st ed. Apress, 2014.

[8] (2015, Sep.) Why hackers are targeting health data. [Online]. Available: http://www.databreachtoday.asia/hackers-are-targeting-health-data-a-7024

[9] (2015, Sep.) Your medical record is worth more to hackers than your credit card. [Online]. Available: http://www.reuters.com/article/2014/09/24/us-cybersecurity-hospitals-idUSKCN0HJ21I20140924

[10] Y. Zhang and C. Poon, "The development of health care data warehouses to support data mining," *Clin Lab Med.*, vol. 28(1), pp. 55–71, 2008.

[11] S. Nugawela, "Data warehousing model for integrating fragmented electronic health records from disparate and heterogeneous clinical data stores," M.Sc. Thesis, Queensland University of Technology, Australia, 2013.

[12] W. Kerr, E. Lau, G. Owens, and A. Treer, "The future of medical diagnostics: large digitized databases," *Yale J Biol Med*, vol. 85, no. 3, pp. 363–377, 2012.

[13] S. I. Khan and A. S. M. L. Hoque, "Towards development of health data warehouse: Bangladesh perspective," in *Proc. 2nd International Conference onElectrical Engineering and Information Communication Technology (ICEEICT)*, May 2015, pp. 1–6.

[14] S. I. Khan and A. Hoque, "Towards development of national health data warehouse for knowledge discovery," in *Intelligent Systems Technologies and Applications*, ser. Advances in Intelligent Systems and Computing. Springer-Verlag, 2016, vol. 385, no. 2, pp. 413–421.

[15] S. I. Khan and A. S. M. L. Hoque, "Development of national health data warehouse for data mining," *Database Systems Journal*, vol. VI, no. 1, pp. 3–13, 2015.

[16] (2015, Jul.) A quiet revolution: Strengthening the routine health information system in bangladesh. [Online]. Available: http://health.bmz.de/good-practices/GHPC/A_Quiet_Revolution/HIS_Bangladesh_long_EN.pdf

[17] (2016, Feb.) Reported breaches of compromised personal records in europe. [Online]. Available: http://cmds.ceu.edu/sites/cmcs.ceu.hu/files/attachment/article/663/databreachesineurope.pdf

[18] IBM and P. Institute, "2015 cost of data breach study: Global analysis," IBM and Ponemon Institute, Research Report, 2015.

[19] P. Institute, "Fifth annual benchmark study on privacy & security of healthcare data," Ponemon Institute, Research Report, 2015.

[20] (2016, Feb.) Breach portal: Notice to the secretary of hhs breach of unsecured protected health information. [Online]. Available: https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf

[21] (2016, Feb.) Hospital pays hackers 17,000 to unlock ehrs frozen in 'ransomware' attack. [Online]. Available: http://www.modernhealthcare.com/article/20160217/NEWS/

160219920/hospital-pays-hackers-17000-to-unlock-ehrs-frozen-in-ransomware

[22] (2016, Feb.) Jackson health: rogue employee suspected of stealing private patient information. [Online]. Available: http://www.miamiherald.com/news/health-care/article59339038.html

[23] (2016, Feb.) 91k patients data compromised in wa healthcare data breach. [Online]. Available: http://healthitsecurity.com/news/91k-patients-data-compromised-in-wa-healthcare-data-breach

[24] (2016, Feb.) Missing drives contained phi on 950k centene customers. [Online]. Available: http://www.scmagazine.com/missing-drives-contained-phi-on-950k-centene-customers/article/467860/

[25] (2015, Sep.) Premera blue cross breach exposes financial, medical records. [Online]. Available: http://krebsonsecurity.com/2015/03/premera-blue-cross-breach-exposes-financial-medical-records/

[26] (2016, Feb.) Anthem hit by massive data breach. [Online]. Available: http://www.healthcareinfosecurity.com/anthem-health-hit-by-massive-data-breach-a-7876

[27] Verizon, "Protected health information data breach report," Verizon, Research Report, 2015.

[28] (2016, Jan.) Lincare ordered to pay 239,800 hipaa privacy penalty. [Online]. Available: http://www.modernhealthcare.com/article/20160209/NEWS/160209856/lincare-ordered-to-pay-239800-hipaa-privacy-penalty

[29] L. Jin, C. Li, and S. Mehrotra, "Efficient record linkage in large data sets," in *Proc. Eighth International Conference on Database Systems for Advanced Applications (DASFAA 2003)*, Mar. 2003, pp. 137–146.

[30] E. Sauleau, J. Paumier, and A. Buemi, "Medical record linkage in health information systems by approximate string matching and

clustering," *BMC Med Inform Decision Making*, vol. 5, pp. 32–44, 2005.

[31] N. K. Abel, P. C. John, L. J. Kathryn *et al.*, "Design and implementation of a privacy preserving electronic health record linkage tool in chicago," *Journal of the American Medical Informatics Association*, pp. 1–9, 2015.

[32] S. I. Khan and A. Hoque, "Privacy and security problems of national health data warehouse: A convenient solution for developing countries," in *Proc. 2nd International Conference on Networking Systems and Security (NSysS)*, Jan. 2016, pp. 157–162.

Shahidul Islam Khan, Abu Sayed Md. Latiful Hoque

Received October 21, 2015

Revised April 5, 2016

Dept. of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka-1000
E-mail: nayeemkh@gmail.com
asmlatifulhoque@cse.buet.ac.bd