Applied Mathematics & Information Sciences
*An International Journal*

http://dx.doi.org/10.12785/amis/080630

# Software Development for Transitions of Graphs from Discrete State into the Continuous State

*Ahmet Koltuksuz and Çağatay Yücel*[*]

Department of Computer Engineering , Yaşar University, Universite Caddesi, No:35-37, Ağaçlı Yol, Bornova, İzmir PK. 35100, Turkey

Received: 31 Oct. 2013, Revised: 29 Jan. 2014, Accepted: 30 Jan. 2014
Published online: 1 Nov. 2014

**Abstract:** Manifolds are suitable differentiable mathematical objects for information to be defined on. By their very definition they are non-Euclidean in the global view but in local scales they resemble Euclidean spaces. This property provides that the contemporary information models can also be defined within the previsioned new models of information models. One of the most basic representations of information is through graphs. They are discrete and highly computable mathematical objects. In this research, the main aim is to investigate methods of embedding this simple piece of information onto manifolds. This research shows that the very fundamental data structures of computer science can be transformed into the continuous spaces and wide area of applications can be engineered such as pattern recognition or anomaly detection. The visualizations of the inspected methods are the evidence of that the graph data can carry new characteristics other than classical properties of graphs such as curvature, locality or multi-dimensionality.

**Keywords:** Information Modelling, Graphs, Manifolds, Laplace-Beltrami Operator, Continuity

## 1 Introduction

The information model is the representation of information in a way that it can be analyzed, measured, processed and transferred. The contemporary information model deals only with the syntactic information, such as frequency of the occurrences of characters, length of words and compression percentage of plain texts. The model was introduced by Claude E. Shannon in his 1947 famous paper A Mathematical Theory of Communication [11].

In this syntactic information model, the definition of information is based on probability theory and statistics. The Shannon Entropy, the most striking concept within this model, is given by the quantification of the expected value of information contained in a message. Shannon's model contains nothing about the semantics of information. For the semantic properties to be modeled, ontology based semi-automatic information retrieval models have been proposed in the literature [12]. These models rely mostly on the human interaction to define the relations between words, in order to derive their meanings.

Information Retrieval (IR) is the process of searching specific information either as text, sound, image, video, data or metadata in a set of documents within a collection. The Vector Space Models (VSMs) have been the standard model for information retrieval since 1975. In this model, some words; which are determined as unique, or some subset of unique words within document collection represents a dimension in space and called as terms. Choosing the terms depends on the application. Each of these documents and queries represents a vector within that multi-dimensional space.

VSM terms are assumed to be orthogonal. This assumption leaves out the semantic relationship between terms. The terms; which represent the coordinate system of the document space, can be related in such a way that, the angles between them are skew-angular instead of being orthogonal. This problem is called The Problem of Dimensionality [1].

Regarding the coordinate system as constant is yet another problem in addition to the problem of dimensionality. The angles between terms can vary depending on the document. This variation among documents leads to the document spaces to be curved because of the varying coordinate system with respect to the document. In this problem, the space that the documents reside may have different angles between coordinates hence resulting a curvilinear document space.

[*] Corresponding author e-mail: cagatay.yucel@yasar.edu.tr

The aforementioned problems lead to the assumption that the structure of information is non-linear, and should be defined on continuous mathematical objects instead of vector spaces. Therefore, the models related to the manifolds are studied in this research. Manifolds are suitable differentiable mathematical objects for information to be defined on. By their very definition they are non-Euclidean in the global view but in local scales they resemble Euclidean spaces. As a consequence, the contemporary models can also be defined within the previsioned new models of information models.

One of the most basic representations of information is through graphs. Graphs are discrete and highly computable. In this study, the main aim is to investigate methods of embedding information onto manifolds using graphs. The methodology is constructed as follows;

- The graph should be constructed from points which are believed to be samples from a manifold, so that the geometry of information is preserved.
- The relation between the properties of the graph and the manifold should be defined.
- And finally, the embedding map should be constructed.

Transition of graphs onto manifolds enables a series of applications such as graph matching and dimensionality reduction to be accomplished using graphs along with the manifold properties. Image, text and sound analysis examples can be found in [10][3] [9]. Though the methods of transition of graphs borrowed from the areas of pattern recognition or manifold learning, the perception of these methods in the process of modeling information and the idea of inspecting information as a curvilinear space is new.

## 2 Mathematical Background

**Laplacian for Graphs Definition 1** The Laplacian can be defined as $L = D - W$:

$$L(u,v) = \begin{cases} d_v - w_{uv} & \text{if } u = v \\ -w_{uv} & \text{if } a_{ij} \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $D$ is the Diagonal Weight Matrix, $W$ is the Weight Matrix, $L$ is the resulting Laplacian Matrix of the graph and their elements are represented by $d$ , $w$ , $L(u,v)$ respectively. The indices u and v are positive integers in the range of $[0,n]$ where $n$ is the number of nodes and $a_{uv}$ defines the elements of the adjacency matrix of the graph.

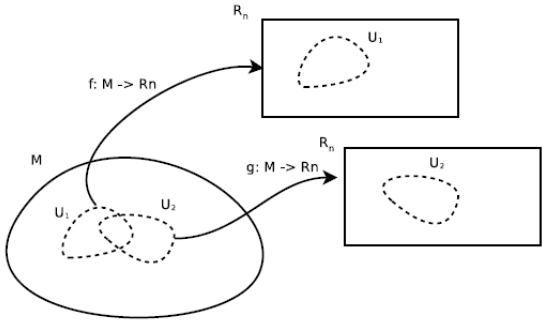



**Fig. 1:** Manifold Definition

### *2.1 Manifolds, Tangent Spaces and Laplace-Beltrami Operator*

**Definition 2** A manifold is a topological space which can be covered by collection of open subsets $O_i$, where $O_i$ is isomorphic to some open subset of $R^n$. This definition is visualized in Figure 1.

The subsets $U_1$ and $U_2$ are mapped onto Euclidean spaces by two maps $f$ and $g$. They are called diffeomorphisms. The definition of diffeomorphism is given below.

**Definition 3** A diffeomorphism between two manifolds is a differentiable map which possesses a differentiable inverse. Also, a smooth map between two manifolds is always continuous.

The first concept in the above definition is being locally Euclidean. The images of charts are Euclidean spaces and since all the charts are consisting of an open set and a map, the chart resembles the Euclidean space of the same dimension. This property is called being locally Euclidean.

The other important property among charts is being smoothly sewn together. The meaning of this property is that diffeomorphisms can be defined between the intersectional parts between the Euclidean spaces that the local parts of the manifold resemble. Figure 2 pictures this property.

### *2.2 Directional Derivatives and Tangent Spaces*

A tangent space at a point p can be imagined as the collection of vectors that is tangent to all the curves passing through p. A derivative definition of manifolds on curves should be given next in order to define the concept of "being tangent on manifolds".

**Definition 4** Let $F$ be the space of all curves through a point $p$ on a manifold. For each differentiable curve $f$ in $F$, there is an operator called *directional derivative* such that:

$$f \to df/d\lambda$$

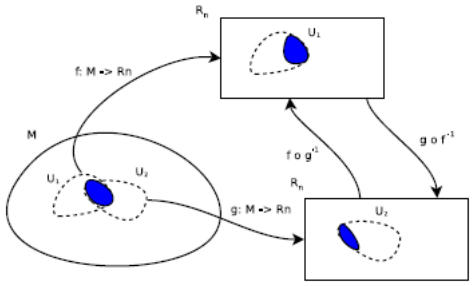



**Fig. 2:** Concept of being smoothly sewn

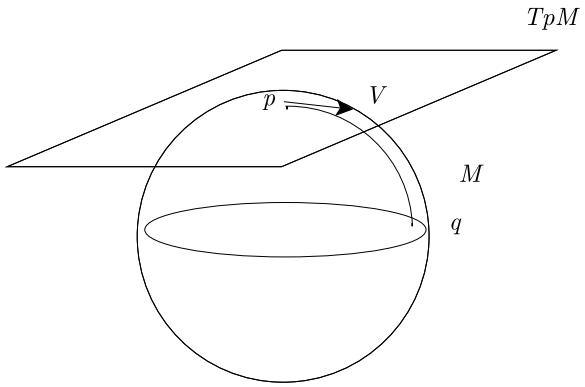



**Fig. 3:** Exponential Map of a vector $v$ at point $p$

where $\lambda$ is the parameter along the curve.

Being differentiable for a curve on a manifold is satisfied when the curve is differentiable at every chart of the manifold. With the definition of a derivative on manifolds, we can claim that a tangent space is the space of directional derivative operators along the curves through $p$. The tangent space definition is as the following [6]:

**Definition 5** Tangent space is a real vector space $\mathbb{R}^n$ tangentially attached to a point $p$ of a differentiable $n$-manifold $M$, denoted by $T_pM$. If $\gamma$ is a curve passing through $p$ then the derivative of $\gamma$ at $p$ is a vector in $T_pM$.

## *2.3 Riemannian Metric and the Metric Tensor*

At every point of a manifold, there is a tangent space that defines the tangent vectors of that point. The tangent space at a point $p$ has the same dimensionality as the manifold.

There are two properties for a manifold to be Riemannian: it should have an inner product defined in every tangent space of the manifold such that one can compute the norm of a vector and the distance between two vectors from that space. The other property is that the inner product should vary smoothly and inner product of two tangent spaces should specify a smooth function on $M$. This inner product property is allowed by the metric tensor [6].

Since the basis vectors of the tangent space can be constructed using the partial derivatives of the manifold at a point $p$, the metric can also be different at every point on the manifold and the metric should vary smoothly from point to point on the manifold as the coordinate system changes. That means precisely, given any open subset $U$ on manifold $M$, at each point $p$ in $U$, the metric tensor assigns a metric $g_{\mu,\vartheta}$ and this assignment is a smooth mapping on $M$. Furthermore, it can be seen as a bilinear operator on vectors $V^{\mu}, U^{\vartheta}$ and also denoted as $g_p(V^{\mu}, U^{\vartheta})$.

## *2.4 Gradient, Exponential Map and Laplace-Beltrami Operator*

**Definition 6** The gradient of a scalar function on $M$ is the vector directed at the greatest rate of change and has magnitude of the greatest rate of change at the point $p$.

$$\text{grad}(f_p) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right)$$

Another definition should be given in order to define Laplace Beltrami Operator which is the main object of study in this research. With the use of the definition of geodesics we can define the exponential map of a vector in a tangent space of a manifold.

**Definition 7** The *exponential map* $Exp_p$ at a point $p$ in $M$ maps the tangent space $T_pM$ into $M$ by sending a vector $v$ in $T_pM$ to the point in $M$ a distance $|v|$ along the geodesic from $p$ in the direction of $v$ [8].

The exponential map takes a vector from the tangent space and maps it onto another point on the manifold using the geodesic along the direction of the vector. Figure 3 depicts the map from the tangent space at $p$ onto the point $q$.

**Definition 8** The Laplace-Beltrami operator is denoted as $\triangle$ and defined in euclidean spaces as

$$\triangle_M f(p) = \sum_i \frac{\partial^2 f(exp_p(v))}{\partial x_i^2}$$

and on any manifold as

$$\triangle_M f(p) = \frac{1}{\sqrt{det(g)}} \cdot \sum_j \frac{\partial}{\partial x^j}\left(\sqrt{det(g)} \cdot \sum_i g^{ij} \cdot \frac{\partial f}{\partial x_i^2}\right)$$

where $f : M \to R$ is a scalar function, $g^{ij}$ is the metric of the manifold.

### 2.5 Convergence of the Laplacian to the Laplace-Beltrami Operator

In this part of the study, the convergence and relation between Laplacian and Laplace-Beltrami Operator is inspected. The foundations of this theorem are given by Belkin and Niyogi in their 2008 paper [4]. The theorems for uniform and random distributions are as follows:

**Theorem 1** Let data points $x_1,\ldots,x_n$ be sampled from a uniform distribution on a manifold $M \subset R^n$. Put $t_n = n^{-\frac{1}{k+2+\alpha}}$, where $\alpha > 0$ and let $f \in C^\infty(M)$. Then the following equation holds:

$$\lim_{n\to\infty} \frac{1}{t(4\pi t)^{\frac{n}{2}}} L_n^{t_n} f(x) = \frac{1}{vol(M)} \triangle_M f(x)$$

where the limit is taken in probability and $vol(M)$ is the volume of the manifold with respect to the canonical measure.

**Theorem 2** Let $P : M \to R$ be a probability distribution function on $M$ according to which data points $x_1,\ldots,x_n$ are drawn in independent and identically distrubuted manner. Then for $t_n = n^{-\frac{1}{k+2+\alpha}}$, $\alpha > 0$, we have

$$\lim_{n\to\infty} \frac{1}{t(4\pi t)^{\frac{n}{2}}} L_n^{t_n} f(x) = \frac{1}{vol(M)} P(x) \triangle_{P^2} f(x)$$

where $\triangle_{P^2}$ is the weighted Laplacian.

## 3 Python Coding of Graph Embedding Methods

In this study, graph embedding methods are coded in the programming language of Python version 2.7. Python language is chosen because of the fast n-dimensional matrix manipulation library NumPy and the scientific library of Python SciPy. The versions of NumPy and SciPy are respectively 1.6.1 and 0.9.0. Open source mathematical software SAGE is used to produce the manifold visualizations by the B-spline method. The version of SAGE used in this study is version 4.8.

### 3.1 Laplacian Eigenmaps

Laplacian Eigenmaps method considers the construction of geometric representation of data on a low dimensional manifold. The geometrical intuition behind this method is inspired by the convention of heat in the nature. This method constructs a natural link between the Laplacian for Graphs and the Laplace Beltrami Operator by the heat equation [3].

In this method, locality of the nodes with respect to their Euclidean distances is preserved. Locality property means that the embedding keeps the local points near on the manifold. The neighborhood information also plays a key role in the construction of the graph from datasets. The graph is constructed by $k$-nearest neighbors $(K-NN)$ or $\in$-neighborhood. In either case the locality is tried to be preserved and the near points are tried to be connected, which ensures the neighborhood information also to be preserved. The algorithm is as follows:

**Algorithm 1** Laplacian Eigenmaps [3]

1.Constructing the adjacency graph using
- $k-NN$ or
- $\in$ – Neighbourhood.

2.After constructing the adjacency graph. The graphs weights should be chosen. Two ways defined in the Laplacian Eigenmaps method. These are:
- Simple minded weight selection:

$$w_{ij} = \begin{cases} 1 & \text{if node i and j are connected} \\ 0 & \text{otherwise} \end{cases}$$

- The heat kernel weight selection, which is:

$$w_{ij} = \begin{cases} e^{-\frac{|x_i - x_j|^2}{4t}} & \text{if node i and j are connected} \\ 0 & \text{otherwise} \end{cases}$$

3.Construct the Graph Laplacian and compute the eigenvalues and eigenvectors for the problem of:

$$L \cdot f = \lambda \cdot D \cdot f \tag{2}$$

Let $f_0, f_1, \ldots, f_{k-1}$ be the solutions of the problem 2. The solutions are ordered according to their eigenvalues:

$$L \cdot f_0 = \lambda_0 \cdot D \cdot f_0$$
$$L \cdot f_1 = \lambda_1 \cdot D \cdot f_1$$
$$\ldots$$
$$L \cdot f_{k-1} = \lambda_{k-1} \cdot D \cdot f_{k-1}$$
$$0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{k-1}$$

The embedding is constructed by omitting the $f_0$ since it is the trivial solution of the problem 2. [3]

### 3.2 Locally Linear Embedding

LLE method is one of the dimensionality reduction methods with a different approach. LLE, instead of estimating pairwise distances, globally reconstructs the embedding using an error function on linear weights. This error function is used to keep local points near in the embeddings. The linear weights are computed as the minimal value of the following error function:

$$\varepsilon(W) = \sum_i |X_i - \sum_j W_{ij} X_j|^2 \tag{3}$$

The weights of the graph from the sample points are constructed by minimizing this least square problem. In this computation, there are two constraints: only the connected points are accounted for the least square problem and sum of all edge weights of each node is always 1. By these two constraints, the constructed graph presents invariant information about the underlying geometry [10].

**Algorithm 2** Locally Linear Embedding [10]

1.For each node in the dataset, the edges are defined by either $k-NN$ or $\in-$ neighbourhood.
2.Each edge given a weight in the interval of $[0,1]$ by minimizing the function $\sum_i |X_i - \sum_j W_{ij} X_j|^2$ such that the sum of all weights of each node is 1.
3.Embedding is computed by taking $k$ lowest eigenvectors of the matrix:

$$E = (I-W)^T (I-W)$$

### *3.3 A Riemannian Approach for Graph Embedding*

In this method the same relationship between Laplacian and Laplace-Beltrami operator is used. However, the edge weights are chosen as sectional curvatures of a manifold with constant curvature. This method uses the properties of Jacobi fields to compute an edge-weight matrix in which the elements are connected by curved geodesics on the manifold between nodes.

In general, manifolds can have rather complex structures than the constant curved ones. However, the approach of this method is the most geometrically intuitive one. Finding a manifold which encapsulates the underlying geometry of information is the main aim of this method. The embedded manifold assumed to be of constant curvature. The curvature is represented by a parameter $K$, such that $K \in \mathbb{R}$. By altering this parameter, one can try to approach the geometry of underlying manifold of information.

The formulation in step 2 of Algorithm 3 is the representation of the geodesics on the manifold with the constant curvature $\kappa$. The function $a(u,v)$ is the Euclidean distance of the two nodes $u$ and $v$. When $\kappa = 0$, that means the space is flat. On that ground, the edge weights are equal to the weights of an Euclidean space. If $\kappa \neq 0$ then the corrections which reflects the diversion from euclidean space is included in the formulation. This corrections are calculated as the Jacobian Field of a geodesic from a manifold of constant curvature [9].

**Algorithm 3** A Riemannian Approach for Graph Embedding [9]

1.For each node in the dataset, the edges are defined by either $k-NN$ or $\in-$ neighbourhood.
2.Each edge given a weight by the function:

$$W_{ij} = \begin{cases} \int_0^1 (a(u,v)^2 + \kappa(sin(\sqrt{\kappa}a(u,v)t)^2))dt & \kappa > 0 \\ \int_0^1 a(u,v)^2 dt & \kappa = 0 \\ \int_0^1 (a(u,v)^2 - \kappa(sinh(\sqrt{-\kappa}a(u,v)t)^2))dt & \kappa < 0 \end{cases} \quad (4)$$

3.The embedding is calculated as the eigenvalues of the Graph Laplacian as explained in the third step of the Algorithm 1.

## 4 Results

- This research is a new approach to the information theory besides being a survey of graph embedding methods. The methods are investigated through a new point of view such that the information theory should be defined on a new definition of information which is continuous and non-Euclidean.

- Graphs are mapped on the manifolds in this research as they are non-Euclidean and continuous by definition and visualizations are done with the B-spline surface visualization method to provide a geometric approach of projections of the discrete data on a continuous surface. These projections show the geometric properties of a piece.

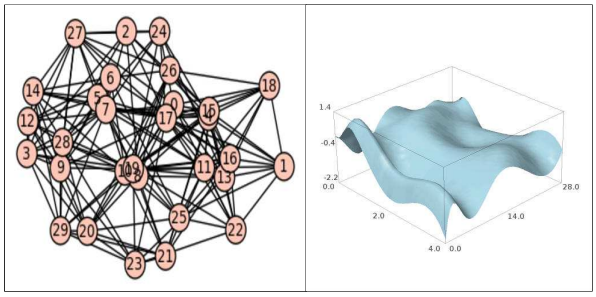


**Fig. 4:** Visualization generated for the Laplacian Eigenmaps Method for a graph of 30 nodes.

- One main result of this research is the revelation of the need of the new definitions for the information. This research shows that the very fundamental data structures of computer science can be transformed into the continuous spaces and wide area of applications can be engineered such as pattern recognition or anomaly detection.

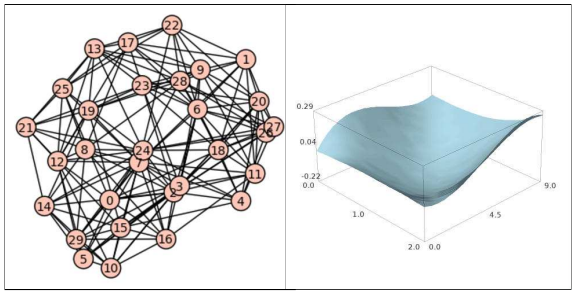


**Fig. 5:** Visualization generated for the LLE Method for a graph of 30 nodes.

- The visualizations show that the graph data can carry new characteristics other than classical properties of graphs such as curvature, locality or multi- dimensionality.These properties vary according to the data and therefore the corresponding embedding space. Our research intuitive is that these graphical properties can create a new variety of research areas when the metrics of the information system are defined, thus leading to new engineering applications in various fields.

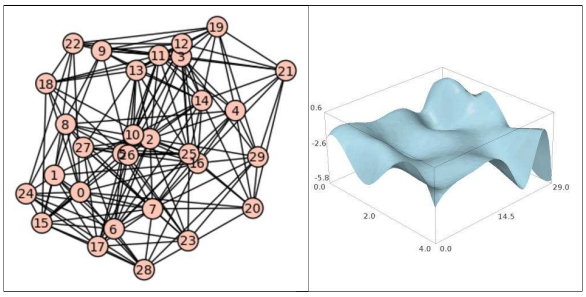


**Fig. 6:** Visualization generated for the Riemannian Approach for a graph of 30 nodes.

## 5 Conclusion

- It is shown that the mapping from discrete domain in the form of a graph into a continuous one; namely a manifold, can be done through Laplace-Beltrami operator. Although this operator has been quite well known and been extensively studied plus utilized by many mathematicians [5], [14], [13], [7], [2], as is been shown by this study, it can be a frame of a reference for the new information model.

- Thus, the point of origin of this research is that the information model should be smooth and nonlinear. To define a new information model, the properties and analogies between discrete and continuous worlds is inspected via this research.

- The link between one of the main data structures of computation and smooth manifolds is investigated. Several methods are evaluated for the purpose of establishing the link in between.

## References


[1] A. Koltuksuz and S. Tekir, Intelligence Analysis Modeling, International Conference on Hybrid Information Technology, EEE Computer Society, 146-151 (2006).

[2] P. Amritanshu, Eigenfunctions of the Laplace-Beltrami Operator on Hyperboloids,Tamkang Journal of Mathematics, 335-339 (2008).

[3] M. Belkin and P. Niyogi, Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, Neural Computation , 1373-1396 (2003)

[4] M. Belkin and P. Niyogi, Towards a theoretical foundation for Laplacian-based manifold methods, J. Comput. Syst. Sci., Academic Press, Inc., 1289-1308 (2008).

[5] M. P. Carmo, Differential Geometry of Curves and Surfaces, Prentice-Hall, (1976).

[6] S.Carroll, Spacetime and Geometry: An Introduction to General Relativity, Benjamin Cummings, (2003).

[7] V. A. Menegatto, Old and New on the Laplace-Beltrami Derivative, Numerical Functional Analysis & Optimization, 309-341 (2011).

[8] F. Morgan, Riemannian Geometry: A Beginners Guide, A K Peters/CRC Press, (2009).

[9] A. Robles-Kelly, A Riemannian approach to graph embedding, Pattern Recognition, 1042-1056 (2007).

[10] S. T. Roweis and L. K. Saul, Nonlinear Dimensionality Reduction by Locally Linear Embedding, Science, 2323-2326 (2000).

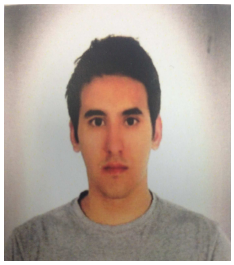
[11] C. E. Shannon, A mathematical theory of communication, Mobile Computing and Communications Review, 3-55 (1947).

[12] D. Vallet, M. Fernandez and P. Castells, An Ontology-Based Information Retrieval Model, The Semantic Web: Research and Applications, Berlin /Heidelberg: Springer, 455-470 (2005).

[13] J. Wu, M. Chi and S. Chen, Convergent discrete Laplace-Beltrami operators over surfaces, CoRR, abs/1004.3486 (2010)

[14] D. L . Zhang, A discrete scheme of Laplace Beltrami operator and its convergence over quadrilateral meshes, Computers & Mathematics with Applications, 1081-1093 (2008).

**Ahmet Koltuksuz** was born in 1961, received his Masters Degree from the Computer Engineering Department of Aegean University with a thesis of Computer Security Principles in 1989. Earned his Ph.D. from the same Institution with a dissertation thesis of Cryptanalytical Measures of Turkey Turkish for Symmetrical Cryptosystems in 1995. And, appointed as an Assistant Professor subsequently. He moved to Izmir Institute of Technology, Department of Computer Engineering in 1996 and became a full-time, tenured Associate Professor within the same institution in 1999.Dr. Koltuksuz had established & run the Information Systems Strategy and Security Laboratory (IS3 Lab) in there. He joined to the department of Computer Engineering of the College of Engineering of Yaar University in September 2009 and is now the head of the department. Dr. Koltuksuz has initiated the Cyber Security Graduate level program effective 2012 fall term, in Yaar University and is currently chairing it. His research interests are Cryptology, Theory of Numbers, Information Theory, Theory of Computation, Operating Systems, Multicore Architectures, Cyberspace Defense & Security, and Open Sources Intelligence Analysis and of Computer Forensics.



**Çağatay Yücel** is a Ph.D. candidate and a Research Assistant at Yasar University, Izmir. He graduated from Computer Engineering from Izmir Institute of Technology at 2009. He got his masters degree in Computer Engineering from Yasar University at 2012. His research interests are Theory of Computation, Computer Security, Cryptography, Theory of Information, Cyber Warfare and Cyber Espionage.