Journal of Mobile Multimedia, Vol. 6, No. 3 (2010) 243–262

ROBUSTNESS OF DYNAMIC SOCIAL NETWORKS

MAYTHAM SAFAR[a]
Computer Engineering Department, Kuwait University, Kuwait
Al-Khaldiya, Kuwait City, Kuwait, Tel:+(965) 24987962
maytham.safar@ku.edu.kw

HISHAM FARAHAT
Computer Engineering Department, Kuwait University, Kuwait
Al-Khaldiya, Kuwait City, Kuwait, Tel:+(965) 24987412
hishamfarahat@gmail.com

KHALED MAHDI
Chemical Engineering Department, Kuwait University, Kuwait
Al-Khaldiya, Kuwait City, Kuwait, Tel:+(965) 24985618
k_mahdi@kuc01.kuniv.edu.kw

Received December 16, 2009
Revised May 17, 2010

The cyclic entropy of a real virtual friendship network provides an insight on the degree of its robustness. Cyclic entropy depends on counting the number of cycles of different sizes in the network, in which a probability distribution function is resulted. Counting the number of cycles in the network is an NP problem. In this work we used a polynomial time approximation algorithm to count the number of cycles in an undirected graph that is based on regression and on a statistical mechanics approach. We used this approximation algorithm to analysis the dynamicity of a virtual social network, E-mail Messages Exchange Network (EMEN) where nodes and edges appear and disappear through time. We analyze the exact and approximated cyclic entropy variation with time as a function of the number of nodes and edges in the network. We further compare the cyclic entropy variation of the network to the traditional degree entropy variation. The purpose is to establish the robustness of the network. In addition, we study the effect of weighed links (number of interactions between users) on the analysis of such graphs. An actual friendship network is found to have cyclic entropy bounded between random and small-world networks models.

Keywords: Email Analysis, Cyclic Entropy, Cycles, Graphs, Directed, Undirected
Communicated by: D. Taniar

1 Introduction

The communication field is one of the largest fields in engineering, and it is growing faster than any other field [1]. Telephones, mobiles, LANs, E-mail [2], chatting [3], peer-to-peer networks and even friends websites [4]. All these facilities and many others were created to facilitate the communications between two or more users and appease the users' sociality [5].

[a]Computer Engineering Department, College of Engineering and Petroleum, Kuwait University, PO Box 5969, Safat 13060, Kuwait

Sociality is the most uniqueness of the human being. Humans usually strive to create relations with others by sharing their thoughts, emotions, and even their actions. Sometimes, it is not necessary even to be a direct interaction between actors to say that there is a social relation between them. It is enough that one of them is acting under the assumption that the others shared the same meanings that cause him to act.

Sociology is the branch of social sciences that considers investigating empirically the social activities of the human being. Its concerns include both micro and macro levels of the human-to-human interactions. In other words, it considers both the face-to-face human interaction and the overall society behavior [6, 7]. After the technological advancement in the communication field and the creation of the Internet and mobiles, the ability to provide insight has grown tremendously. Fortunately, the ability to study the human society and answer the questions mentioned has grown. Thanks to the facilities mentioned above that usually store information that is enough to be used for modeling these communications. By utilizing this information, it becomes possible to answer this question in accuracy that makes it worthy to be considered. This stored information acquired the researchers' interests and attracted them to use it in their researches. And that is what created a new term called "Social Network," a graph that represents each actor in the community as a vertex, and the relations between actors as an edge. Social networks are how any community is modeled. The social network model helps the study of the community behavior and thus leveraging social network to researchers' demands [8, 4, 9].

Numerous applications of social networks modeling are found in the literature. Some researchers tried to use email inbox as a source to develop a social network and use it to fight spam messages [10]. Others tried to use data stored in banks, phone records, vehicle sales, surveillance reports and registration records to create a social network and to analyze this network for the purpose of fighting criminal organizations [11]. Some other researches tried to use social networks to represent web-communities to analyze the World Wide Web. Other applications include data model, compression methods, indexing and query operators were suggested in [8], [12], and [13] respectively to analyze social networks. With the increasing population of the world, the importance of modeling social networks and analyzing their robustness increases.

An essential characteristic of any network is its resilience to failures or attacks, or what is known as the robustness of a network [8]. The definition of a robust network is rather debatable. One interpretation of a robust network assumes that social links connecting people together can experience dynamic changes, as is the case with many friendship networks such as Facebook, Hi5, etc. Individuals can easily delete a friend or add a new one, with and without constraints. Other networks, however, have rigid links that are not allowed to experience changes with time such in strong family network. Entropy of a network is proven to be a quantitative measure of its robustness. Therefore, the maximization of a network's entropy is equivalent to the optimization of its robustness. Albert et al. [8] describe the effect of a network's heterogeneity on its degree of tolerance against either random node failures or intentional attacks. The three models of social networks are analyzed and compared: Scale-free (SF), Random Networks (RN) and Small-World (SW). Scale-free networks, which include social networks, were found to display a high degree of robustness against random failures but great vulnerability against targeted attacks. Many researchers, who mostly used percolation

theory to study the resilience of different complex network topologies, further investigated the study and analysis of resilience in complex networks. Methods based on Percolation Theory focus on analyzing a threshold value, which represents the number of nodes that must be removed from a network before it disconnects into smaller, separate networks. Conversely, [9] studied the robustness of scale-free networks to random failures using entropy of the degree distribution in the network, hence the level of its heterogeneity. An optimal design of a robust network was achieved through the maximization of its entropy, following a nonlinear mixed integer programming approach.

Entropy is a very important characteristic that has been used to determine the degree of robustness in social networks [14, 15, 16, 17]. Entropy of a network is related to the probability of finding the network in a given state. For a system of moving molecules, the state is obviously the positions and the momentum of each molecule at a given instant. For a system of magnets, the state is defined through the magnets directed north or south. The entropy of a specific network shape was investigated before in [18], where the entropy of a Lattice network was studied. While this Lattice was a theoritcal one, our work studies the entropy of actual social networks .

Most of the previous mentioned studies characterize social networks using degree distribution, clustering coefficient, average length and average degree [19], or assume that the social network is static. In our work, social networks evolve over time, driven by the exchanged data between its members and the appearance/disappearance of the members from the network and we analyze cycles instead of degrees. Recently in [3, 14] we proposed a model to compute a statistical mechanical property, the cyclic entropy of the network as a measure of the degree of network robustness. Such property was based on counting the number of cycles (circuits) of certain length existing in the network. The problem of finding and counting circuits in large graphs has been of interest to researchers lately due to its challenging complexity; has been known to be an NP-Complete problem. Exhaustive enumeration, even by smart algorithms proposed in earlier research, is restricted to small graphs as the number of circuits grows exponentially with the size of the graph [3, 20]. Therefore, it is believed that it is unlikely to find a precise and efficient algorithm for counting circuits. Finding a method of approximation to this problem is the alternative. In this work we illustrate a method of approximation that has been suggested to estimate, in polynomial time, the number of circuits in a graph as a function of their length. The algorithm is based on a work done by [21]. In this work, we extend our previous study in [2] by analyzing the behavior of a social network created by email exchange, however, we looked at a longer period of time and compared it to the traditional degree distribution analysis. We study the behavior of the social networks when taking into consideration the weights of links between the nodes. Those weights represent the degreenumber of email messages exchanged between users in the social network.

2 Related Work and Background

With the increasing population of the world, the importance of modeling social networks increases. Numerous applications of social networks modeling exists in the literature. In [22] they tried to use data stored in banks, phone records, vehicle sales, surveillance reports and registration records to create a social network and to analyze this network to fight criminal organizations. Some other researches tried to use social network to represent web-communities

to analyze the World Wide Web. Other applications include data model, compression methods, indexing and query operators were suggested in [8, 12, 13] respectively to analyze social network.

Work in [23] use graphs derived from e-mail to define the organizational structure of a corporation. Another work in [24] detected telephone fraud by comparing the social behavior of new telephone accounts to that of previously tagged fraudulent accounts. In [10] they tried to use a mail inbox as a source to develop a social network and asses the network with the objective of fighting spam messages. In [25] a study was conducted about the spread of computer viruses via e-mail messages. The address books form a directed social network of connections between individuals over which the virus spreads. They investigated the structure of this network and discussed its implications for the prevention of computer virus epidemics.

The research in [26] addressed the problem of correctly relating aliases that belong to the same entity. Their network was constructed from email data mined from the Internet. Links in the network represent web pages on which two email addresses are collocated. The work in [27] analyzed e-mail social network analysis for the detection of security policy violations on computer systems. They assume that the properties of social networks are computationally feasible to evaluate, and in fact can be determined in linear time. In addition, the authors were not able to predict a universal social structure which can be exploited for finding all the violations. The study in [28] analyzed a dynamic social network in which interactions between individuals are inferred from time-stamped e-mail headers recorded over one year. They discovered that the evolution of such a network is related to both the network topology and the application area in which the network is embedded. However, they assumed that global perturbations of such networks are absence, and they used average network properties. Another work in [29] automatically extracted social hierarchies from Enron corporations e-mail data to analyze and catalog patterns of communications between entities to rank relationships. It assumes that the organization is dynamic and its structure changes over time.

Network robustness is also a vital property that was considered in many literatures [19, 30, 17]. A dynamic system is said to be robust if it is resilient to attacks and random failures. There are several types of threats that a robust network must be secured from. Random vertex removal, an intentional attack to vertices, a network fragmentation and any other event that causes a reduction in the network information-carrying ability can be considered as a threat. In [31] they experimentally found that a scale-free network shows a good resilience to random failures. The heterogeneity of the network degree distribution dictates the chance of randomly attacking a crucial vertex. Depending on this remark, criteria to characterize the complex networks robustness by measuring its heterogeneity were suggested in [17]. They tried to use the principle of entropy to calculate how much the network degree distribution is unbalanced and thus the network heterogeneity. However, this method is considering only the random failures. It was proven also that unbalanced degree distribution causes in contrast very low intentional attack survivability [31]. Removing the small partial of vertices that have the most connections will cause total network destruction. In principle, entropy is roughly related to the degree of disorder in the system and how much it is stable. The entropy is not known to be strongly related to the heterogeneity. Calculating it statistically suggests defining the microstates that define the system configuration. And this deludes that there is a strong relation between the entropy and heterogeneity. This argument motivates us to

suggest a new technique to evaluate social network robustness. It is true that the robustness measurement can be done by calculating its entropy.

The study in [31] describe the effect of a network's heterogeneity on its degree of tolerance against either random node failures or intentional attacks. The three models of social networks are analyzed and compared: Scale-free (SF), Random Networks (RN) and Small-World (SW). Scale-free networks, which include social networks, were found to display a high degree of robustness against random failures but great vulnerability against targeted attacks. The study and analysis of resilience in complex networks was further investigated by many researchers, who mostly used percolation theory to study the resilience of different complex network topologies. Methods based on Percolation Theory focus on analyzing the threshold value pc, which represents the number of nodes that must be removed from a network before it disconnects into smaller, separate networks. Conversely, [17] studied the robustness of scale-free networks to random failures using entropy of the degree distribution in the network, hence the level of its heterogeneity. An optimal design of a robust network was achieved through the maximization of its entropy, following a nonlinear mixed integer programming approach. The authors in [14] propose a universal distribution function form based the degree of loops or cycles existing in the network instead of the degree of links in the network. The network configuration state was thus defined as the degree of cycles within the network rather than the common definition of the network state as the degree of links associated with the actors in the social network. This new distribution form was found applicable to all types of social networks (scale-free, small world, and random networks). The same definition of the system state was used in [15] on a fully connected social network for the purpose of finding the maximum entropy value, hence identifying the equilibrium state of the social network, the state of maximum entropy. In other words, finding the point where the system is most stable.

3 E-mail Messages Exchange Networks (EMEN)

Email messages exchanging is one of the most common way of communication between people. It is one of the easiest ways to exchange information between distant individuals in a very fast way and with minimum setup cost. Most organizations consider the email as a formal communication between the employees and other organizations. E-mail messages are exchanged on a daily basis between individuals, corporate and educational departments. The E-mail logs over a period of time of a user can generate predictable patterns in the social network that can be quantified using graph theory.

The scope of the current work is to analysis a dynamic social network, that is a network with no restrictions in changing the relations between the actors. In other words, nodes and edges appear and disappear through time. An example of such a network is a popular virtual social network in the Internet known as Email Messages Exchange Network (EMEN). Based on social networks definition, EMEN is considered as a social network where its actors are individuals and relations are email messages exchanged between them. The action of exchanging emails between two individuals is considered as a relation between them.

In this paper, we detail the use of e-mail analysis for the detection of cycles and computing the cyclic entropy of such a network. We construct the network from data obtained from E-mail account of a user, where the nodes are considered as the users in the To:, From:, CC:,

and BCC: fields of an email message. The links are weighted links that represent the number of email messages exchanged between two different users. Then, we analyze the cyclic entropy variation with time with the aim of studying the robustness of such a network. Techniques similar to those discussed in this paper can be applied to spam filtering, computer virus spread prediction, and e-mail manageability.

4 Entropy of Cycles (Loops)

Loops were one of the major concerns in social network field. In [32], they mentioned that the loops (cycles) can be considered as the major aspect that can separate the graph to sub-graphs or components. Other researches (e.g., [33] and [34]) proved that there is a strong relation between the structural balance of a social network and the loops in the network. Our previous works presented in [3, 14, 20] define the relation between network robustness and the entropy. They defined the entropy as "the degree of disorder in the system". From statistical mechanics, the entropy can be calculated from a given probability distribution $P(k)$ of the system in state k:

$$S = -\sum_{k} P(k) ln P(k). \tag{1}$$

We stated that a robust network (unconstrained and dynamic network) has low entropy, and a static rigid network has large entropy. We proposed a new concept of computing the entropy, the network cycles' distribution is used instead of the well known degree based entropy. The cyclic entropy characterizes the network more accurately than the degree entropy.

Based on the definition of entropy in Eq.(1), we need to evaluate from real data the distribution function that characterizes EMEN social network. As stated earlier, entropy of a network is related to the probability of finding the network in a given state. In social network, there are several choices that define the state of the network; one is the number of social links associated with a social actor, known as degree. This definition is commonly used by almost all researchers. In [3, 20] we showed that the characterization of social networks through the degree leads to different non-universal forms of distribution, and that there is no universality class reported. Here we propose a universal distribution form that is applicable for all social networks by using a different definition of the state of network. We define the state as the degree of loops or cycles exiting the social network. Then we find the distribution function of such loops. We analyze the probability of an actor receiving the same E-mail message dispatched by him/her again?

5 Network Modeling

5.1 Social network graph representation

The analysis of a network system needs the network to be modeled mathematically as a graph [35]. The graph theory has been used to analyze and compute the robustness of the EMEN social network. Here we proposed three models that generate three undirected graphs. The graphs are different in what the edge will represent. In the first model (Directional Binary Email Exchange DBEE) , there will be an undirected edge between two individuals if either one sends an email to the other. In the second model (Compulsory(Mandatory) Email Exchange CEE), an undirected edge between two individuals means that both individuals

have sent an email to the other. While in the third model (Weighted Email Exchange WEE), the edge has another property which is the weight of the edge.

5.1.1 Directional Binary Email Exchange (DBEE)

Let $G(V, E)$ be an undirected graph representation of the network, where V is the set of email addresses (individuals) that have sent exist in the EMEN. And E is the set of edges, where e represented as the tuple (u, v, c) which means that u and v have communicated by email (either u sent an email to v or vice versa) , c times in a window of time. To illustrate how the EMEN can be modeled as a graph, assume the situation stated in Table 1.

Table 1. Example of small data gathered from the log file.

ID	From	To	CC	BCC
Email1	1	2	5	-
Email2	1	2	3	4
Email3	2	3	-	5
Email4	4	5	-	3
Email5	5	2	-	1
Email6	2	1	-	-

An example of DBEE model is shown in Figure 1, where the nodes represent the the email accounts. The labels on the edges represent the number of email messages. An example of a cycle is $< 1, 5, 4, 3, 1 >$, which has a length of 4.

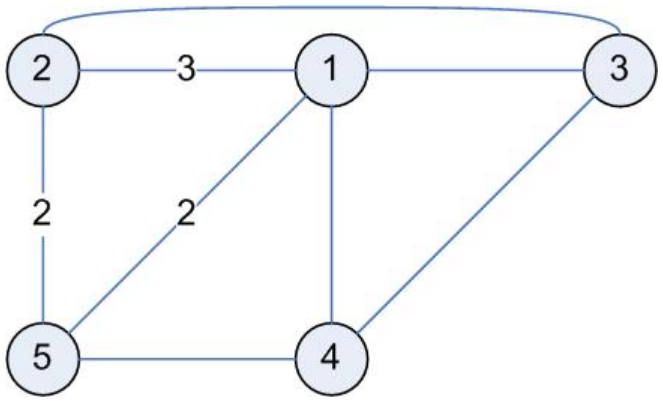

Fig. 1. DBEE model Example.

5.1.2 Compulsory(Mandatory) Email Exchange (CEE)

Let $G(V, E)$ be an undirected graph, where V is the same set as DBEE model. And E is the set of edges, where e represented as the tuple (u, v, c) where there exist two(or more) email messages, at least one from u to v and one from v to u, and

$$c = \left\lfloor \frac{\text{Total number of emails between } u \text{ and } v}{2} \right\rfloor . \tag{2}$$

An example of CEE model is shown in Figure 2 (un-weighted) , that is based on the

example in Figure 1. The nodes represent the individuals. The only cycle is $< 1, 2, 5, 1 >$, which has a length of 3.

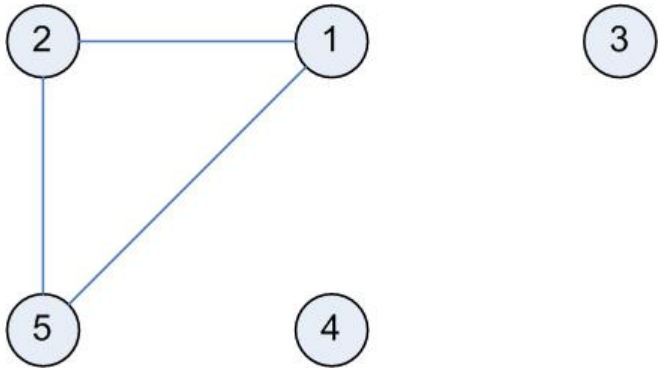

Fig. 2. CEE model Example.

5.1.3 Weighted Email Exchange (WEE)

This model is a variation of the CEE model. It adds weights to the edges to construct a weighted undirected graph. Let E is the set of edges, where e represented as the tuple (u, v, c, w) where u , v and c have the same definition of the CEE model. And

$$w = \frac{c}{\text{Maximum c of all edges}}. \tag{3}$$

In this model the cycle has two properties, length and weight. Based on the previous example the cycle has a length 3 and a weight of $\frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1$, check Figure 3.

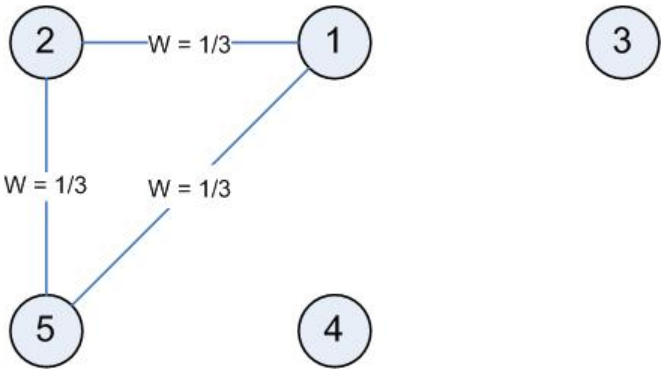

Fig. 3. WEE model Example.

6 Cycles Computation

To analyze our social network, we will only consider cycles as the main parameter that characterizes our graphs. A simple cycle is a sequence of nodes (path) $< v_0, v_1, .., v_k >$ if $k > 3, v_0 = v_k$ and $v_1, v_2, .., v_k$ are distinct. Two cycles are distinct if one is not a cyclic permutation of the other [36].

In [3, 20], we computed the cycles distribution using exact and approximate algorithms that are based on the works in [36] and [21], respectively. In both algorithms, we assumed that we have as an input a undirected graph $G(V, E)$ where V is a set of vertices and E is a set of ordered pairs called edges, which is represented as adjacency list AG. We assume that those graphs have: 1) No self-loops, i.e. no edges of type (v, v), 2) No multi-edges , i.e. no two edges has the same source and destination, 3) G is strongly connected graph, 4) The vertices are numbered with IDs from 1 to n.

6.1 Exact cycles distribution algorithm

With this algorithm we are interested in finding the number of cycles for each possible cycle length. We developed a java program to find all the simple cycles in a graph that is based on an algorithm created by Johnson [36]. The algorithm is based on backtracking technique. It starts with node s which is the vertex with the least ID, and begins to romanlist all cycles that passes through s. This is done by building a simple path starting from s using a stack to save the vertices. Whenever s is encountered again a cycle is created and printed. In Addition to that any vertex is currently in a path (stored in the stack) is being blocked so it cannot be added again to the stack. When a node is finished (the algorithm passes through all of its edges), it is being popped from the stack and unblocked for future use. After enumerating all the cycles with s as common node, the algorithm removes s from the graph G and starts again the process with the second least vertex. These steps are done until G has two nodes. Since our graph is undirected, the equation

$$C_{un} = \frac{C_{d\sigma} - e}{2}. \tag{4}$$

is used to convert from directed cycles to undirected cycles. Where $C_{d\sigma}$ is the total number of cycles in the directed version of the original undirected graph (i.e. replace each undirected edge with two directed edges in opposite directions). C_{un} is the total number of cycles in the undirected graph and e is the total number of edges in the graph.

The pseudo code for this algorithm is shown in Algorithm 1. For extra details on the exact algorithm, please refer to [3]. The complexity of Johnson's algorithm is $O((n + e)c)$ where n is the number of nodes, e is the number of edges and c is the number if circuits in the graph.

6.2 Approximate cycles distribution algorithm

The proposed work depends on the cycle's distribution of networks. To get the cycles distribution, we need to find the number of cycles for each possible cycle length in the network. We developed a java program in that is based on the backtracking algorithm [21]. However, due to the complexity of such computation (NP-Hard) [37], we developed a new approximation algorithm for counting cycles in a network [20, 38]. The approximation is based on a statistical mechanics [39] approach designed by [21]. It uses the Belief Propagation equations [40, 41] and an approximation method to approximate the statistical mechanics model and find the cycles distribution. Two methods can be used as approximation algorithms, Monte Carlo simulation and Bethe approximation. Bethe is used here because of the well-known correspondence between both Bethe and Belief Propagation.

The graphs in this algorithm are represented as adjacency matrices. The input to the algorithm is an undirected graph, and the output is the cycle's distribution of the graph

Alg. 1 CircuitFinding(G).

Input: $G = (V, E)$ Graph
Integer List arrays $A_k[n], B[n]$,
Boolean array *blocked*$[n]$, Integer s

```
1begin
2       empty stack
3       s ← 1
4       while (s < n) do
5            Ak ← adjacency structure of strong component K with least vertex in subgraph
       of G induced by {s, s + 1, ..., n}
6            if (A_k ≠ φ) then
7                 s ← least vertex in V_k
8                 for (i ∈ V_k) do
9                      blocked(i) ← false
10                     B(i) ← φ
11                end
12                dummy ← CIRCUIT(s)
13                s ← s + 1
14           else s ← n
15      end
16end
```

Procedure **CIRCUIT(Integer** v**)**

```
1begin
2       f ← false
3       stack v
4       blocked[v] ← true
5       for (w ∈ A_k[v]) do
6            if (w = s) then
7                 output circuit composed of stack followed by s
8                 f ← true
9            end
10           else if !blocked(w) then
11                if CIRCUIT(w) then f ← true
12      end
13      if f then UNBLOCK(v)
14      else for (w ∈ A_k[v]) do
15           if v ∉ B[w] then put v on B[w]
16      unstack v
17      return f
18end
```

Procedure **UNBLOCK(Integer** u**)**

```
1begin
2       blocked (u)← false
3       for (w ∈ B[u]) do
4            delete w from B[u]
5            if blocked(w) then UNBLOCK(w)
6       end
7end
```

(number of cycles as a function of their size). The algorithm starts by reducing the graph. All leaf nodes (nodes with degree 1 or 0) are removed from the graph. Each edge of the graph is initialized with a random positive value $y^{(0)}$. Each edge is iterated from its initial value until convergence reaching to a fixed value of y^*. Convergence is determined according to some accuracy level. To guarantee the convergence of the algorithm, we restricted $|y^{T+1} - y^T| \leq 0.001$ to be less than or equal to 0.001. The value y represents the probability that the edge is present in a cycle c. The y value can be calculated using the following equation:

$$y_{i\to j}^{T+1} = \frac{u \sum\limits_{m\in\beta_{i-j}} y_{m\to i}^{T}}{1 + 0.5u^2 \sum\limits_{m,n\in\beta_{i-j}\ m\neq n} y_{m\to i}^{T} y_{n\to i}^{T}} \tag{5}$$

where u is a positive real value. Then from all y's two values are calculated; C_L and

$$L = \sum_{(i,j)\in E} \frac{u\, y_{i\to j}^{*}\, y_{j\to i}^{*}}{1 + u\, y_{i\to j}^{*}\, y_{j\to i}^{*}} \tag{6}$$

$$\begin{aligned} R &= \frac{1}{N} \sum_{i\in V} ln \left(1 + 0.5u^2 \sum_{\substack{m,n\in\beta_{i-j} \\ m\neq n}} y_{m\to i}^{*}\, y_{n\to i}^{*} \right) \\ &- \frac{1}{N} \sum_{(i,j)\in E} ln\left(1 + u\, y_{i\to j}^{*}\, y_{j\to i}^{*}\right) \\ &- \frac{L\, ln(u)}{N} \end{aligned} \tag{7}$$

$$C_L = e^{RN} \tag{8}$$

where

- β_i is the set of neighbors of node i.
- β_{i-j} is the set of neighbors of i except a neighbor j.
- N is the number of nodes in the graph.
- C_L is the number of cycles of size L.

Refer to [21] for further details on the above equations.

The procedure explained above is repeated starting from an initial value of $u = u_0$ to $u = u_{max}$. Where u_0 and u_{max} are greater than 0. At each iteration step, a new distribution point (L, C_L) is produced. The iteration step for u is 0.0001 at the early stages of the algorithm. This value is not fixed. It will be changed when $L_{new} - L_{old} < 0.001$ (i.e. the progress in L is slow). If this condition is satisfied, u will be increased by 10%. As noticed from the equations above, the output at each step (L, C_L) depends on u. At specific stages of the iteration (when u gets large), many iterations are wasted giving nearly the same point. To avoid this condition, a jump in u is made.

This algorithm yields a plot of (L, C_L) points. To extract the needed distribution points (3 to n), we use Gaussian formulation based on the work done in [42], the distribution relates

the number of cycles to cycle length if network parameters (a,b,c) are known or found by non-linear regression. The equation

$$y = a \cdot e^{-(\frac{x-b}{c})^2} \tag{9}$$

(where x is the cycle length L and y is cycles count C_L) models the cycles distributions of a graph. We used Equation 9 to fit the curve and find the function that represent the distribution.

The pseudo code for this algorithm is shown in Algorithm 2. This algorithm has a running time that is growing polynomially with the graph size and logarithmically with the required accuracy. Since the algorithm uses the adjacency matrix to represent the graph, the space complexity is $O(2n^2)$.

Alg. 2 ApproximateAlgo(G).

Input: $G = (V, E)$ Undirected graph
Output: $A[1..n]$ Array of size n which contains the cycle count of each length

```
1   begin
2       A[1] ← A[2] ← 0
3       Reduce G by removing all leaves nodes
4       U ← 0
5       Points ← null set of points
6       while (L < n)
7       begin
8           y ←random number between 0 and 10 ( do for all edges)
9           while (|y_new − y| < 0.0001 for all edges)
10          begin
11              Calculate y_new using equation 5 for all edges twice
12              y ← y_new
13          end
14          L ←equation 6
15          C_L ← equation 8
16          Add (L, C_L) to points
17          u ← u + 0.0001
18          if ( |L_new − L_old| < 0.001)
19              u ← u ∗ 1.1
20      end
21      f(x) = doRegrssion(L,C)
22      for (i = 3 → n)
23      begin
24          C_i = f(i)
25          A[i] ← C_i
26      end
27  end
```

7 Experiments and Results

7.1 Data Extraction

To study and analyze the Email Messages Exchange social network, we have extracted a log of email account messages during a period of 18 months since the date it started. The log contains over 354 unique individuals and 2258 email messages. Figure 4 shows a snapshot of the extracted network at some instant time. The extraction phase was programmed with two languages:

(i) Visual Basic: To access the Outlook and extract the log.
(ii) Java: To analyze the log and give statistics.

7.2 Experiments

We conducted three sets experiments to characterize Email messages exchange social networks using cycles. In the experiments we took bi-monthly snapshots of the network to study the network evolution with time. In all the experiments, we aim to calculate the cyclic entropy which requires counting the cycles. The approximate algorithm is used in the first experiment. The exact computation is used in the second and third experiments since the graph sizes are small. We further compare our cyclic entropy to a more traditional degree entropy that is used by most works in the literature. We have used a regular Intel Core Duo based laptop with 2GB of RAM.

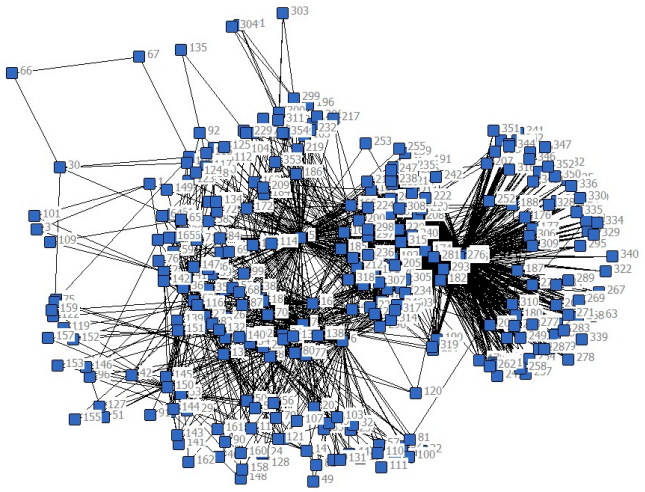

Fig. 4. A snapshot of the used network.

7.2.1 Influence of unitary email exchange on entropy

In this set of experiments, we consider two individuals have a relation (edge between them) if either one sent an email message to the other (i.e., used DBEE model). We first show how the network evolves with time during the 18 months. Figures 5 and 6 show the number of the nodes and edges in the network are increasing logarithmically with time. The algorithms that are being used to compute the number of cycles in a graph have a very high time complexity,

especially for graphs with large number of nodes, edges and hence cycles which is our case. DBEE model network is very large to be computed in a finite time. Hence, for this experiment, the approximate algorithm was used to compute the entropy of the network. Figures 7 and 8 draw the cyclic entropy and degree entropy evolution, respectively, of DBEE model network during the 18 months; we can notice that the entropy of the network is also logarithmatically increasing with time as the size of the network increases (especially the number of nodes and not edges). The figures show that the cyclic entropy is precisely describing the network behavior as the degree entropy. However, as we proved in our previous works, that the cyclic entropy can further precisely describe the type of the underlying network.

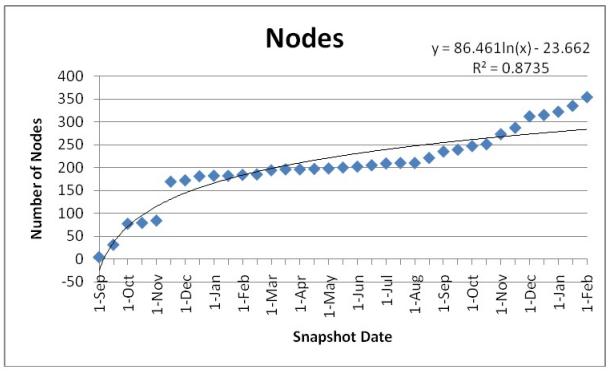

Fig. 5. DBEE model Number of Nodes vs Time.

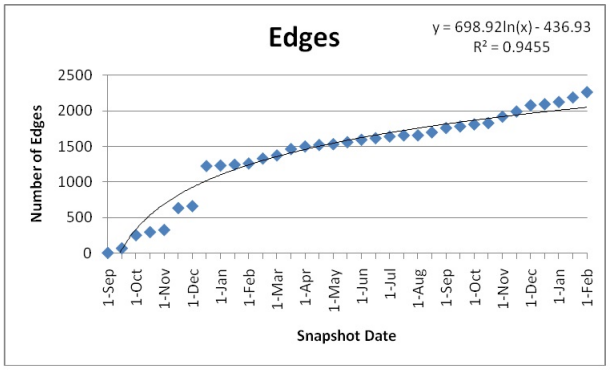

Fig. 6. DBEE model, Number of Edges vs Time.

7.2.2 Influence of binary email exchange on entropy

In this experiment we used the same log file and same snapshots used in first set of experiments to study the network entropy change with time. However, for this experiment we consider two individuals have a relation (edge between them) if both have sent an email message to each other (i.e., used CEE model). Figures 9 and 10 show the linear increase in the number of nodes and edges with time. Since the graph size is small, we applied the exact entropy algorithm to compute the entropy of this network. Figures 11 and 12 illustrate the cyclic and

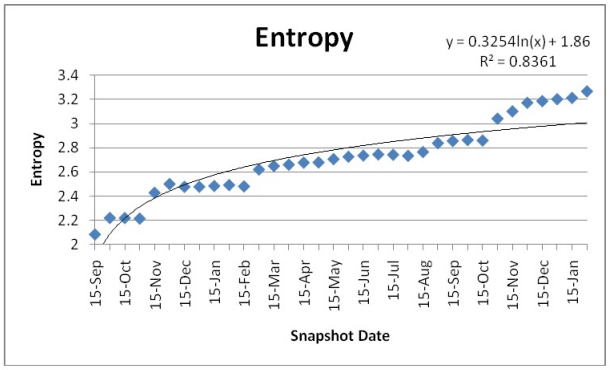

Fig. 7. DBEE model, Cyclic Entropy vs Time.

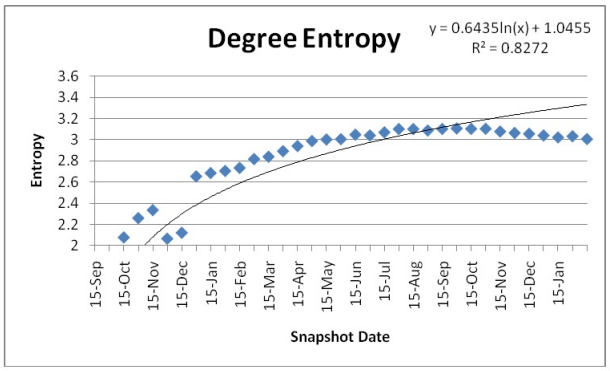

Fig. 8. DBEE model, Degree Entropy vs Time.

degree entropy evolvement, respectively, of the network, that exhibits a linear increase with time. It is also clear that this increase is more related to the number of nodes and not the edges. In Table 2 we show a comparison of both DBEE and CEE models in terms of the growth rate of the number of nodes, edges, and entropy with time.

7.2.3 Influence of weighted links on entropy

This experiment is conducted to compare CEE and WEE models, which is to compare unweighted and weighted graphs. Figure 13 shows the entropy evolution with time for both cases. It is obvious from the graph that the weights of the edges have no effect on the evolution of the entropy with time. This means that the increasing number of nodes and edges have an effect on the entropy, however, the number of interactions between the same users has a lower (if none) affect on the entropy. Only the connection between those two nodes has an effect on the entropy, and not the frequency of their interactions.

8 Conclusion

Robustness of a social network implicitly assumes that the network is resilient to random failures and attacks. Counting cycles in social networks is an NP problem. Hence, in this work we proposed a polynomial time approximate algorithm to count the number of cycles in an

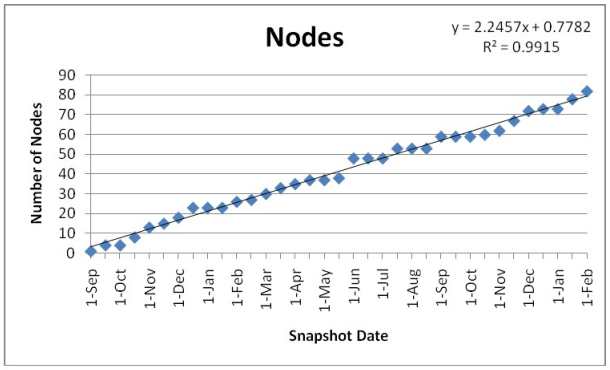

Fig. 9. CEE model, Number of Nodes vs Time.

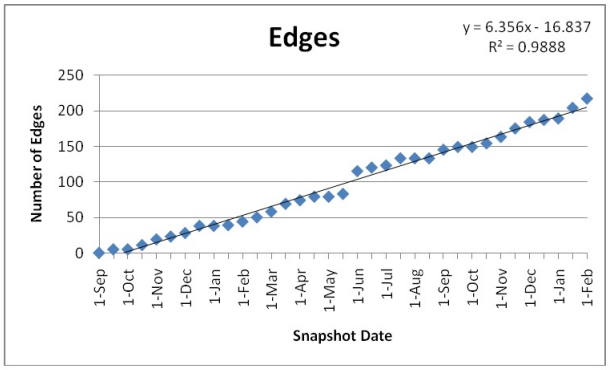

Fig. 10. CEE model, Number of Edges vs Time.

Table 2. Rate of network growth in DBEE and CEE models.

	DBEE Model	CEE Model
Node	$87/n$	2.246
Links	$699/n$	6.356
Entropy	$0.326/n$	0.0657
Degree Entropy	$0.644/n$	0.045
Order	$O(1/n)$	$O(t)$

undirected graph that is based on regression. The approximate algorithm is based on a statistical mechanics approach that uses a Bethe approximation technique and iterations of the Belief Propagation equations. The approximate algorithm is effective in approximating the probability distribution of the cycles in a fraction of the time taken by the exact algorithm. It also achieves order of magnitudes of improvement in running time. Cyclic entropy is used to characterize the dynamic of EMEN using the probability distribution of the cycles computed by the approximate algorithm. In addition, we compute degree entropy and compare it to cyclic entropy. Three cases are examined in this work. The most significant results are of the first two experiments. Both clearly show no equilibrium can be reached as time passes. The Email Exchange network will never reach equilibrium, in other words, their entropy are

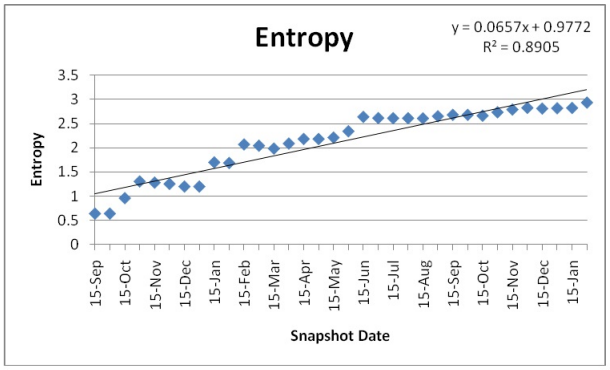

Fig. 11. CEE model, Cyclic Entropy vs Time.

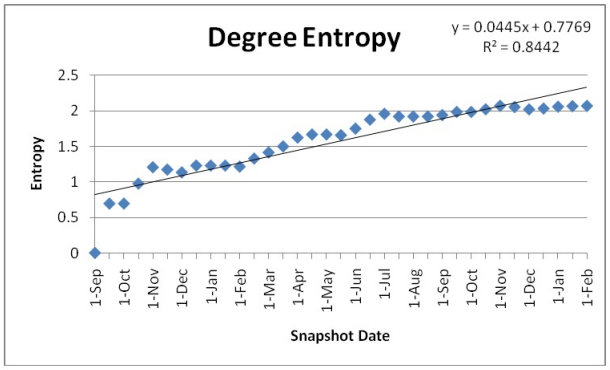

Fig. 12. CEE model, Degree Entropy vs Time.

monotonic functions with time. Looking at the influence of optional email exchange, we observe logarithmic trend of both cyclic and degree entropies with time , $O(ln(t))$. Whereas, the influence of compulsory binary email exchange shows a linear trend with time, $O(t)$. The former case suggests that a slower rate of increase and a slower growth, that slows down as the number of nodes and links increase. The later case has a constant rate of nodes and links increase and entropy as well; such a network can be uncontrollable as time passes. The result of compulsory binary exchange can be utilized to slow down the growth of email exchange growth simply by forcing binary interactions. At time passes, there is an increase probability that more people come and join the email exchange network, hence, the more entropy is added to the system. If the new members are ignored or less contacted by other and more binary exchange among the present members exists, the network's growth should slow down. These trivial observations are nicely presented with the use of cyclic entropy of the network. Moreover, the third experiment has shown that having more interactions between two users in a network does not have an effect on the entropy of the system. While, the number of the users and the links between them (at least one interaction) that effect the entropy of the network.

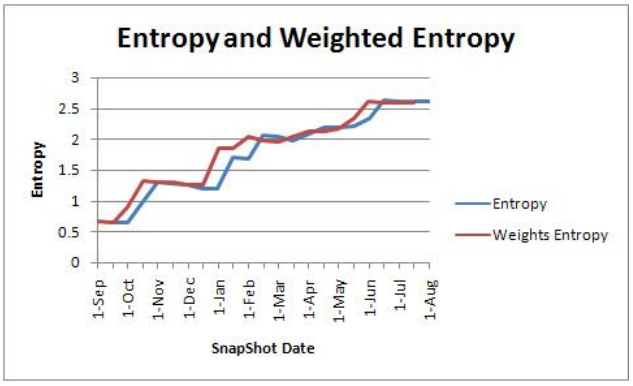

Fig. 13. CEE model vs WEE model.

In the search for a design of most robust network, we propose the use of cycle distribution instead of degree distribution for many reasons. Degree distribution is one dimensional hence it suggests little information on the nature of the network. As for cycles distribution, it is a two dimensional problem that provides more elaborate information about the network. In previous work, the authors showed that cycle distribution provide solid and unique evidence of the type of actual social network through the analysis of its cyclic entropy. In addition, cycles distribution is found to have one universal mathematical representation where degree distribution is mathematically specific depending on the type of social network. Such uniformity of cycles distribution allows better characterization of social networks. The calculation of entropy using the cyclic method will be compared in the future to other methods of calculations. Furthermore, cyclic calculation of entropy is a novel concept that can be explored in further details by considering several types of social networks. Considering the network evolution models, nodal attributes models or exponential random graph models is a must due to their generality and their ability of dynamic network representing. However, applying our methodology in characterizing the network needs an efficient algorithm to count the cycles for such dynamic models. As we have shown in this work, the most important part in the cyclic entropy calculation is counting the cycles. In our future studies, we will study the impact of the increased triads on the network cyclic entropy. We will investigate how to use the triads' information to approximate the computation of the cyclic entropy of a network.

References

1. R.Y. Shtykh, and Q. Jin (2008), *Harnessing user contributions and dynamic profiling to better satisfy individual information search needs*, International Journal of Web and Grid Services, 4(1): 63-79.
2. M. Safar, H. Farahat, and K. Mahdi (2009), *Analysis of Dynamic Social Network: E-mail Massages Exchange Network*, Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services (iiWAS).
3. K. Mahdi, H. Farahat, and M. Safar (2008), *Temporal evolution of social networks in paltalk*, Proceedings of the 10th International Conference on Information Integration and Web-based Applications and Services (iiWAS).
4. M. Safar, and H. B. Ghaith (2006), *Friends Network*, IADIS International Conference WWW/Internet, Murcia, Spain.

5. S. Izumi, K. Yamanaka, Y. Tokairin, H. Takahashi, , and N. Shiratori (2009), *Ubiquitous supervisory system based on social contexts using ontology*, Mobile Information Systems 5(2): 141-163.
6. G. Boella, L. Torre, and S. Villata (2009), *Analyzing Cooperation in Iterative Social Network Design*, Journal of Universal Computer Science, Vol. 15, No. 13, pp. 2676-2700.
7. J.J. Macionis (2007), *Sociology*, Prentice Hall, 12 edition.
8. A. Bagchi, A. Bandyopadhyay, and S. Mitra (2006), *Design of a data model for social network applications*. Journal of Database Management.
9. S. A. J. Shirazi (2006), *Social Networking: Orkut, Facebook, and Gather*, Blogcritics.
10. P. Boykin and V. Roychowdhury (2005), *Leveraging social networks to fight spam*, In IEEE Computer Society Press, Computer, 38(4):61–68.
11. J. Xu, and H. Chen (2005), *Criminal Network Analysis and Visualization*, Communications of the ACM, vol. 48, pp. 100 - 107.
12. T. Bhanu, S. M. A. Bagchi, and A. Bandyopadhyay (2006), *Pre-processing and path normalization of a web graph used as a social network*, Special Issue on Web Information Retrieval of JDIM.
13. S. Mitra, A. Bagchi, and A. Bandyopadhyay (2006), *Complex queries on web graph representing a social network*, 1st International Conference on Digital Information Management, pages 430–435.
14. K. Mahdi, M. Safar, and I. Sorkhoh (2008) *Entropy of robust social networks*, Proceedings of IADIS International e-Society Conference.
15. M. Safar, K. Mahdi, and I. Sorkhoh (2008), *Maximum Entropy of Fully Connected Social Network*, Proceedings of the International Association for Development of the Information Society (IADIS) International Conference on Web Based Communities.
16. I. Sorkhoh, M. Safar, and K. Mahdi (2008), *Classification of Social Networks*, Proceedings of the International Association for Development of the Information Society (IADIS) WWW/Internet Conference.
17. B. Wang, H. Tang, C. Guo, and Z. Xiu (2005), *Entropy optimization of scale-free networks robustness to random failures*, Physica A, vol. 363, Issue 2, pp. 591-596.
18. D. Simovici (2007), *Metric-Entropy Pairs on Lattices*, Journal of Universal Computer Science, 13(11): 1767-1778.
19. R. Albert and A.-L. Barabasi (2002), *Statistical mechanics of complex networks*, Reviews of Modern Physics, 74.
20. K. Mahdi, M. Safar, and H. Farahat (2009), *Analysis of temporal evolution of social networks*, The Journal of Mobile Multimedia (JMM), 5(4):333–350.
21. E. Marinari, R. Monasson, and G. Semerjian (2006), *An algorithm for counting circuits: application to real world and random graphs*, Europhysics Letters.
22. C. Hsinchun and X. Jennifer (2005), *Criminal network analysis and visualization* Communications of ACM, pages 100–107.
23. J. Tyler, D. Wilkinson, and B. Huberman. *Email as spectroscopy: Automated discovery of community structure within organizations*, Communities and Technologies, 6(3):81–96.
24. C. Cortes and D. Pregibon. *Communities of interest*, Intelligent Data Analysis, 6(3):211–219.
25. M. Newman, S. Forrest, and J. Balthrop (2002), *Email networks and the spread of computer viruses*, Physics Reviews, 66(3).
26. R. Holzer, B. Malin, and L. Sweeney (2005), *Email alias detection using social network analysis*, International Conference on Knowledge Discovery and Data Mining archive, Proceedings of the 3rd international workshop on Link discovery.
27. R. Rowe, G. Creamer, S. Hershkop, and S. Stolfo (2007), *Automated social hierarchy detection through email network analysis*, International Conference on Knowledge Discovery and Data Mining archive.
28. G. Kossinets and D. J. Watts (2006), *Empirical analysis of an evolving social network* Science, 311(5757):88–90.
29. G. Creamer, R. Rowe, S. Hershkop, and S. Stolfo (2009), *Segmentation and automated social hierarchy detection through email network analysis*, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Advances in Web Mining and Web Usage Analysis.

30. L. d. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas (2007), *Characterization of Complex Networks: A survey of measurements*, Advances in Physics, vol. 56, pp. 167 - 242.
31. R. Albert, H. Jeong, and A.-L. Barabasi (2000), *Error and attack tolerance of complex networks*, Nature, vol. 406, pp. 378-382.
32. J. Scott (2000), *Social Network Analysis: a Handbook*, Sage Publication Ltd, 2nd edition.
33. W. Nooy, A. Mrvar, and V. Batagelj (2005), *Exploratory Social Network Analysis with Pajek (Structural Analysis in the Social Science)*. Cambridge University Press, England.
34. S. Wasserman and K. Faust (1994), *Social Network Analysis: Methods and Applications*. Cambridge University Press, England.
35. G. D. Marco and L. Barolli (2007), *On some current results of graph theory for ad-hoc networks*, Journal of Mobile Multimedia (JMM), 2(4):15–33.
36. D. B. Johnson, (1975) *Finding all the elementary circuits of a directed graph*, SIAM Journal on Computing, pages 77–84.
37. K. Mahdi, M. Safar, I. Sarkhoh, and A. Kassem (2009), *Cycle-Based versus Degree-based Classification of Social Networks* The Journal of Digital Information Management (JDIM), Volume 7, No. 6, pp. 383-389.
38. M. Safar, K. Mahdi, H. Farahat, S. Albehairy, and A. Kassem (2010), *Approximate Cycles Count in Undirected Graphs* The Journal of Digital Information Management (JDIM), (in press).
39. A. Hobson (1971), *Concepts in Statistical Mechanics*, Routledge, 1 edition.
40. P. D. L. Rios, S. Lise, and A. Pelizzola (2001), *Bethe approximation for self-interacting lattice trees*, Europhysics Letters, 53:176–182.
41. O. Shental, P. H. Siegel, J. K. Wolf, D. Bickson, and D. Dolev (2008), *Gaussian belief propagation solver for systems of linear equations*, The IEEE International Symposium on Information Theory, Pages: 1863-1867.
42. M. Safar, K. Mahdi, and A. Kassim (2009), *Universal cycles distribution function of social networks*, The First International Conference on Networked Digital Technologies (NDT).
43. V. Bhatnagar, S. Kaur, and L. Mignet (2009), *A Parameterized Framework for Clustering Streams*, International Journal of Data Warehousing and Mining, 5(1): 36-56.
44. V. Nikulin (2008), *Classification of Imbalanced Data with Random sets and Mean-Variance Filtering*, International Journal of Data Warehousing and Mining, 4(2): 63-78.
45. J. Luo, and X. Ni (2009), *A clustering analysis and agent-based trust model in a grid environment supporting virtual organisations*, International Journal of Web and Grid Services, 5(1): 3-16.