Selection of Most Informative Components in Risk Mitigation Analysis of Supply Networks: An Information-Gain Approach

Avi Herbon, Eugene Levner, Sharon Hovav, and Lin Shaopei

***Abstract*—The concept of a supply network generalizes supply chains, web, multi-media communication networks, and complex technological projects. A problem investigated in this paper is to minimize the economic loss caused by failures and other undesired events in the network. This is done in two steps: first, we find most informative network components and, then, select the optimal set of risk-preventing activities.**

***Index Terms*—Risk analysis, supply networks, most informative components, information gain, entropy .**

I. INTRODUCTION

A *supply network* is a set of sites or facilities or other related entities connected by transportation links whose function is the processing or procurement of raw material or information products, transformation of the latter products into intermediate and finished (material or information) products, transportation and distribution of the finished products to the customers (see Fig. 1).

This concept generalizes the original concept of a supply chain, where mainly material flows are arranged in linear or tree-type network structures. Other notable examples of the supply networks are the web, multi-media communication networks, and large-scale technological projects ([10], [9], and [7]). The definition of a supply chain given by [13], that combines several known definitions, also nicely defines the supply network:

"The management of material, information and financial flows through a network of organizations (i.e., suppliers, manufacturers, logistics providers, wholesalers/ distributors, retailers) that aims to produce and deliver products or services for the consumers. It includes the coordination and collaboration of processes and activities across different functions such as marketing, sales, production, product design, procurement, logistics, finance, and information technology within the network of organizations" ([13], p. 453).

An important task in managing supply networks is to simplify the network structure aiming to minimize the network dimension and size (which, in turn, minimizes the data storage and simplifies computations) without a meaningful harm to the network goals and customer requirements. A major challenge in this respect is to select most meaningful and informative network components. A specific problem investigated in this paper is to minimize economic loss within the network caused by different failures and undesirable events in the network (e.g. an unexpected increase in market prices, equipment breakdowns, shortages in stock, etc). This is done in two steps: First, we find most informative network components and, then, select an optimal set of risk-preventing strategies.

Manuscript received March 8, 2012; revised April 30, 2012

Avi Herbon and Sharon Hovav are with the Department of Management, Bar Ilan University, Ramat Gan, Israel (e-mail:avher@bezeqint.net; sharonh@clalit.org.il).

Eugene Levner is with the School of Economics, Ashkelon Academic College, Ashkelon Israel (e-mail:levner@hit.ac.il)

Lin Shaopei is with the Shanghai Jiao Tong University, School of Mechanical Engineering Shanghai, China (e-mail:splin@sjtu.edu.cn).

We propose an information-theoretic approach to decreasing the size of the supply network based on measuring the information impact (*information gain*) of sub-systems. This is an informational impact of each node in the network which permits us to select the "most important" components. We measure the information gain of each sub-network with the help of different information measures, such as the entropy, Gini index, and the twoing function.

When selecting the optimal set of risk preventing strategies we use the found information gain of each component in combination with the cost and economic impact of each risk preventing strategy. The optimization knapsack-type problem will model the selection process, in which a portfolio (a set) of selected strategies will provide the maximal economic impact (the maximum decrease of economic losses) with a given budget being taken into account.

In the next section, we define risks related to the economic loss in the supply network caused by different failures and other undesirable events. In Section II we define the information gain measures widely used in information theory and data mining; they will be used as tools for solving the problem of minimizing the economic losses introduced in Section III. Section IV provides a case study related to a healthcare supply chain. Section V concludes the paper.

According to [12] (as well as to a common sense), risk is "a situation in which it is possible but not certain that some undesirable event will occur". There is a plethora of other definitions. Most of the risk evaluation approaches from the management point of view include the probability of an undesirable event and the impact and/or severity of how a disruption affects the flow of products, money and information across organizations in a supply network ([6]).

Following [9], we will estimate the risk of economic loss in a supply network by considering two measures: (i) probability of a failure or other unpleasant event, and (ii) impact of the failure (risk event) mainly expressed as an expected loss value in a monetary form.

II. RISKS AND INFORMATION GAINS IN NETWORKS

The supply network is assumed to be consisting of subnetworks some of which are elementary components whereas some other are subsystems composed of several components. The problem analysis starts with the design of a *decision tree T* which is, in fact, a simplified model (or, in other words, a re-configuration) of the considered supply network.

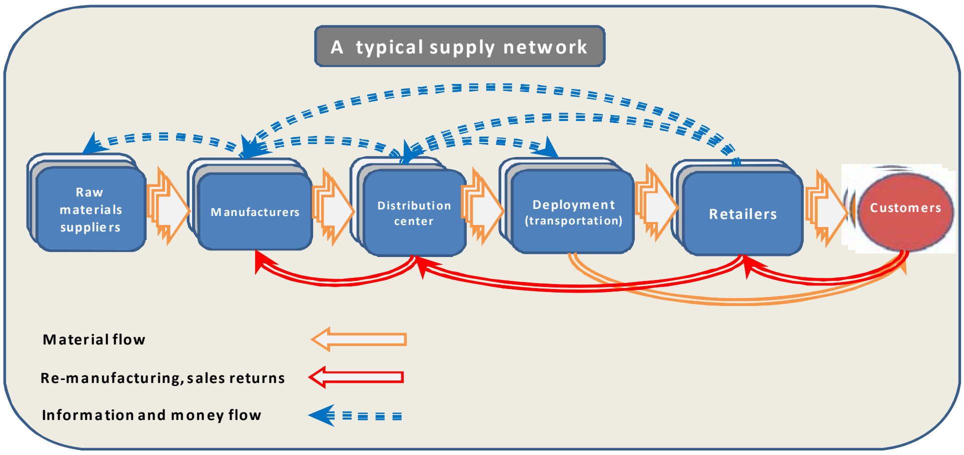

Fig. 1. A typical supply network.

The main idea is that the tree T represents a hierarchy of the supply network; it is constructed sequentially layer after layer, in a top-down fashion. Each node, denoted as v, represents a sub-network J, while child nodes of v are components of the sub-network J. The most top node (a "root") of the tree contains a small number of children nodes. Higher layers of the tree have a much smaller number of nodes than the lower layers.

For instance, a tree root for the network in Fig.2 below may have five nodes corresponding to the main blocks depicted in the figure. It is worth noticing that since the tree T is constructed in a top-down fashion there is no need to break-down the sub-networks in the tree into excessively small parts. Speaking differently, a network designer or a decision maker can stop the process as far as the tree size reaches a pre-specified limit assigned a priori.

Like [2], we are concerned with technological organizational and informational sources of risks. We treat them as attributes assigned to different components of the supply network. In what follows, we call them *risk factors*. . They can be either binary or numerical or ordinal. For simplicity, we consider below only binary ones.

A key concept in our approach for defining the risk factors is an event *list*; this is a list of situations (events) occurring in the supply network (during a certain period) each of which has led or did not led to the total failure of the subsystem considered, and, in turn, to a considerable economic loss. The event list is a table whose rows are events and columns are risk factors.

Each event list is formed for a separate sub-network. Each row represents a situation that has happened at a certain date or during a certain period at the sub-network under consideration. The row components, except for the last one, are the factor values; we use symbol f as a factor index, F is the number of factors, and F+1 is the row length. The last component in a row depicts an event outcome: it is either 1, which means that the subsystem is out-of-order (or, in other words, the situation is *risky*), or 0 which means that the subsystem in the considered situation works well (*not-risky*). In the former case we say that the corresponding event is "of class 1", in the latter case – "of class 0".

The value of the f-th risk factor in any row, say in row r, is 1 if this factor has occurred in the event described by the row r, and 0 otherwise. To illustrate this concept, let us consider a healthcare supply network described in more detail in Section IV. One of the risk factor considered is "*a mistake in forecasting the required demand*". It may happen that for a certain event (say, the event which is represented by row 1 in the event list) this risk factor takes place whereas for some other event (e.g., the event which is represented by row 2) this factor does not occur. In the former case the value of this factor in row 1 is 1, in the latter case it is 0 in row 2. Notice that the outcome of each of the rows considered may be either 1 or 0 depending on the combination of all factors in the row. More than that, even all the factors in a row may be 1 while the outcome of this row maybe 0 (that is, *not-risky*). One can observe that, in fact, the event list is similar to the *training set* in data mining.

Consider a fixed node v and the event list corresponding to this node. As far as the collection of records in the event list is known, a factor f (f = 1,…, F) is called a *risk-driver in a row*, if the factor value in this row is 1, and the row outcome is 1, as well. (Speaking informally, if a factor is a risk driver in a row, it can be a reason for a failure happened in the row). Notice that the fact that a factor f is the risk driver does not claim that the factor causes the risky event but rather it indicates at such a possibility. For each f, we can compute $N_v(f)$ – the total number of those rows in the event list corresponding to node (sub-network) v in tree T in which f is the risk driver. We compute the *relative frequency* $p_v(f|1)$ of cases where factor f (f = 1,…, F) is the risk-driver for the entire list as follows:

$$p_v(f/1) = \frac{N_v(f)}{\sum_{f=1}^{F} N_v(f)} \tag{1}$$

Then the *risk-related node informativeness* (later on simply called the *informativeness*) is defined as the amount of information I contained in the node.

For our aim, I can be any information measure employed in information theory and data mining (e.g., the Shannon entropy, index Gini, classification error, or the twoing function).

One can easily observe that $\sum_f p_v(f|1)=1$. Then the *Shannon entropy* at node v is defined as follows:

$$I(v) = Entropy\ (v) = - \sum_f p_v(f|1) \log p_v(f|1). \tag{2}$$

If I is the *Gini index* at node v then

$$I(v) = Gini\,(v) = 1 - \textstyle\sum_f p_v(1|f). \quad (3)$$

If I is the *classification error* at node v then

$$I(v) = error(v) = 1 - \max_f p_v(1|f). \quad (4)$$

If I is the *twoing function* then

$$I(v) = TF(v) = (w(L)w(R)/4) \textstyle\sum_f | p_v(1|L) - p_v(1|R) |, \quad (5)$$

where L and R stands, respectively, for the left and right daughter node of v .

The informativeness I being defined, the *information gain* (a "goodness") of sub-network v, G (v) is found as follows:

$$G\,(v) = \Sigma_i\, w_i\, I(\text{child\_node\_}v_i) - I(\text{parent\_node } v)\,, \quad (6)$$

where w_i is the weight of the child node i defined as follows:

$$w_i = n(v_i)/n_v. \quad (7)$$

Here n_v is the number of rows in the event list corresponding to parent node v, and $n(v_i)$ is the number of number of rows in the event list corresponding to the child node v_i (we assume that event list gathered for any parent node is the union of the event lists of the children nodes).

When value of G (v) is small there is no sense to break a parent sub-system into a set of children sub-systems whereas if $G(v)$ is big such division is meaningful.

It is impractical and unnecessary to construct the entire decision tree T of the supply chain which may have, in its full form, thousands nodes. The constructing process goes from top down, layer after layer, with the information gain of each node being computed during this process. A threshold value h_0 for the information gain is assigned by a decision maker in advance. Then the nodes whose information gains are smaller than h_0 are discarded and thus, the tree size can be essentially reduced from the initial design T to a new tree, denoted by T'.

III. Selection of Risk-Preventing Activities

After the supply network size is reduced, risk-preventing strategies are defined in each node of the network.

Assume that we have K types of different risks) (for example, technological, informative, organizational, etc.) and S types of risk-mitigating strategies. For each strategy s, $s = 1, 2, \ldots, S$ the following parameters are assumed to be given: strategy cost C_{sk} if strategy s applied for mitigating risk of type k; expected decrease in loss D_{sk} if strategy s is applied for mitigating risk of type k, and the total budget B for carrying-out all risk-preventive and risk mitigating activities.

Then the problem of optimal selection of risk-preventing and risk-mitigating activities aimed to maximize the sum of prevented losses D_{sk} (and, therefore, minimizing the expected economic loss) can be formulated as the knapsack problem.

A. The Multi-Dimensional Knapsack Problem

Let us introduce the following notation:

- s - strategy (activity) index
- k - risk type index
- D_k - expected decrease in loss due to applying all activities for mitigating risk of type k
- ω_k - threshold (lower bound) for the required decrease in loss caused by risk of type k;

The multi-dimensional knapsack problem is formulated as follows:

Maximize expected decrease in loss

$$D = \sum_{k=1}^{K}\sum_{s=1}^{S} x_{sk}\, D_{sx}$$

subject to

$$D_k = \sum_{s=1}^{S} x_{sk} D_{sk} \geq \omega_k\,, \quad k = 1,\ldots,K$$

$$\sum_{k=1}^{K}\sum_{s=1}^{S} C_{sx}\, x_{sk} \leq B$$

x_{sk} = 1 if strategy s is selected for mitigating risk k,
= 0 otherwise.

The first constraint requires the minimum level of risk preventive activities to be carried out; the second one is a budget constraint.

For solving this problem, one can apply any of the standard solution methods developed in literature for the knapsack problem (see[11], and[8]). In particular, in our study we have used a fast greedy heuristics for prioritizing the variables based on ranking decrease-to-cost ratios. Some numerical results related to a real-life problem are given in the next section.

IV. Case Study: The Supply Chain *Clalit*

Large-scale medical supply chains constitute a special class of supply networks. They are characterized by significant social benefits, high monetary value, and intolerance of failures which result in medical item unavailability.

To illustrate the potential for the network size reduction and risk mitigation in this study, a special supply network has been considered. It is a large scale healthcare supply chain for the health maintenance organization called *Clalit*.

The *Clalit* Health Maintenance Organization (HMO), hereafter, *Clalit*, is the second-largest HMO in the world and the largest health organization in Israel. Its structure is presented in Fig.2. *Clalit*'s services are provided to a population of 3.8 million users through a network of 1,300 clinics, 14 hospitals, and 650 pharmacies, as well as hundreds of institutes and laboratories nationwide. *Clalit* employs about 32,000 workers. In this study, *Clalit* is modeled by a supply network whose main sub-networks are:

- Raw material suppliers
- About 350 manufacturers/finished product suppliers
- Three distribution centers
- 14 hospitals, 650 pharmacies and 2,400 clinics
- 3.6 million end customers.

To clarify the presentation, the numerical example is simplified. An initial hierarchical tree T with 30 nodes has been constructed top-down. After the nodes with small information gains have been discarded, twenty most informative nodes remain. For each of them, possible risk mitigating strategies have been found out. Then the set of most effective risk mitigating strategies (activities) has been found by solving the corresponding knapsack problem on the obtained reduced network. For each activity we examined the parameters that were indicated in Section III. The model included a budget limit that made it applicable in practice.

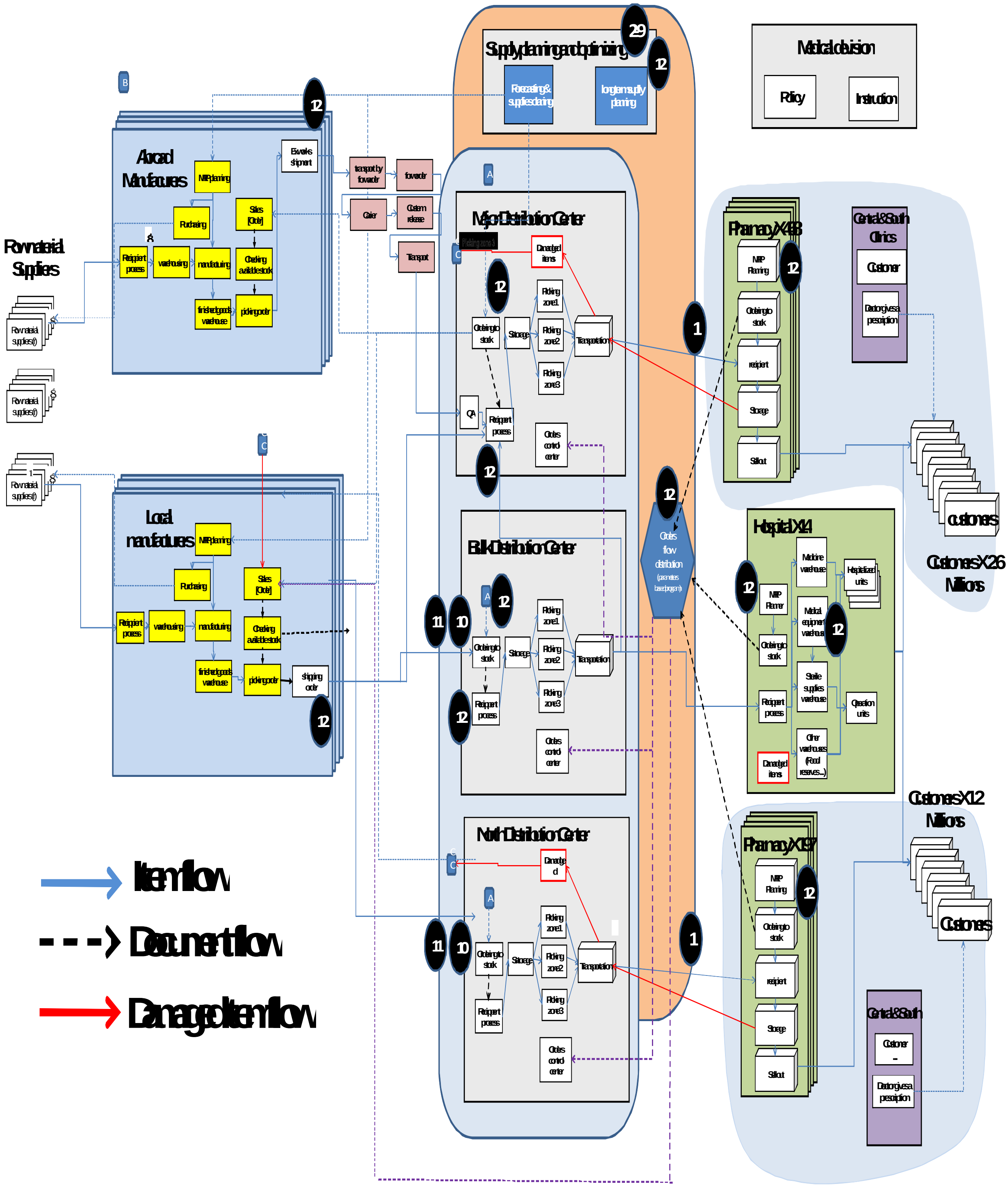

Fig. 2. The healthcare supply network Clalit. Locations of the most important activities are marked by black points.

A. Description of Activities

Activity 1: Setting the optimum period of replenishment (for distribution centers and pharmacies) based on research and recommendations conducted by [5].

Activity 2 – 9: Eight activities for improving forecasting processes related to specific item profiles, for high monetary volume items.

Activities 2 and 3: Planning of new and out-of-date medicines in terms of item lifecycle.

Activities 4 and 5: Removal of peaks and dips in demand forecasts to reduce uncertainty.

Activities 6 and 7: Noise detection and filtering noises as a cleaning process for improving the forecasting.

Activities 8 and 9: Analysis and optimization of items' lifecycle.

Activity 10: Using pharmacies' and hospitals' orders in queue as part of the replenishment order calculation.

TABLE I: INPUT DATA FOR THE CASE STUDY

Activity Number	Activity	Improvement per month [$]	Standard deviation of improvement [$]	Improvement investment cost per month [$]	Savings [$]
Activity 1	Setting periodic review time	50,676	7,601	-21,622	29,054
Activity 2	Phase in/out 'high volume high variety'	7,813	391	-1,028	6,785
Activity 3	Phase in/out 'high volume low variety'	2,622	131	-730	1,892
Activity 4	Peaks/dips removal 'high volume high variety'	7,814	391	-514	7,300
Activity 5	Peaks/dips removal 'high volume low variety'	2,622	131	-912	1,709
Activity 6	Noise detection 'high volume high variety'	7,814	391	-1,028	6,786
Activity 7	Noise detection 'high volume low variety'	2,622	131	-1,635	987
Activity 8	Life cycle management 'high volume high variety'	7,811	391	-2,056	5,755
Activity 9	Life cycle management 'high volume low variety'	2,622	131	-3,269	-648
Activity 10	Extra backlog picking lines	8,516	852	-7,559	957
Activity 11	2 days of stock are invested in safety stock to backup stdv in lead time	8,415	841	-5,405	3,009
Activity 12	Fail because of inaccurate system parameters 2%	5,740	1,148	-946	4,794
		115,083		-46,704	68,380

Activity 11: Increasing safety stock to support lead time fluctuations; analysis of investments in stock versus reduction of backlogs.

Activity 12: Preventing organizational and administrative bureaucracy in replenishment processes. About 2% of replenishment processes was discovered to fail due to inaccurate administration.

The contribution and investment required by each activity are described in Table 1. The optimal solution to the knapsack problem contains 12 activities indicated in Fig. 2. The example shows that by a monthly investment of 170,000$, the expected total decrease of economic loss is 233,400$ per month, so the net savings are about 60,000$.

V. CONCLUSION

The paper presents a framework for the information-gain based selection of most informative components in supply chains and reducing losses due to failures and undesirable events. We are aware that there exist other ways of measuring the node informativeness in graphs and networks (see, e.g.., [1], [3], and [4]). However, a theoretical and practical comparison of these approaches falls out of the scope of this paper

The suggested approach can be extended to treat different types of input data and different classes of risks under uncertainty. In particular, the analysis of stochastic and fuzzy measures for risk assessment and node informativeness as well as the design of practical solution methods for corresponding stochastic and fuzzy knapsack problems are major directions for future research.

REFERENCES

[1] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, 1998, 30 (1-7), pp. 107-117.

[2] M. Cruz and M. L. Pinedo, "Total quality management and operational risk in the service industries," *Tutorials in Operations Research,* chapter 7, 154-168, INFORMS, 2008.

[3] K. Kireyev, "Semantic-based estimation of term informativeness," *Human Language Technologies*: Proceedings of the 2009 Annual Conference of the North American Chapter of the ACL, Boulder, Colorado, June 2009, pp. 530–538.

[4] G. Kasneci, S. Elbassuoni, and G. Weikum, "MING: mining informative entity relationship subgraphs," The 18th ACM Conference on Information and Knowledge Management, CIKM'09, November 2009, Hong Kong, ACM Press, 2009, pp. .1653-1656.

[5] K. Kogan, Y. Perlman, and S. Hovav, "Equilibrium replenishment in a supply chain with a single distributor and multiple retailers," *IMA Journal of Management Mathematics*, 2008, 20, 395-409.

[6] K. Kogan and C..S. Tapiero, *Supply Chain Games. Operations Management and Risk Valuation*, Springer, 2007.

[7] M. Laumanns and E. Lefeber, "Robust optimal control of material flows in demand-driven supply networks," *Physica*, 2006, 363(1) pp. 24-31.

[8] E. Levner and S. Vleduts, Discrete Optimization, Chapter 8 in K.-H. Elster (ed.), *Modern Mathematical Methods of Optimization*, Akademie Verlag, Berlin, 1993, pp. 304-318.

[9] E. Levner and J.-M. Proth, "Strategic management of ecosystems: A supply chain perspective," in E.Levner, I. Linkov and J.-M. Proth (eds), *Strategic Management of Ecosystems*, Springer, 2005, pp. 95-107.

[10] S. P. Lin, "Fuzzy modeling of investment decisions in engineering projects," *Uncertainty Modeling and Analysis in Civil Engineering*, CRC Press, 1998, pp. 167-188.

[11] S. Martello and P. Toth. *Knapsack Problems. Algorithms and Computer Implementations*. Wiley, 1990.

[12] SEP: *Stanford Encyclopedia of Philosophy*, E.N. Zalta (ed), Risk. Summer 2007 Edition, Available on http://plato.stanford.edu/entities/risk

[13] C. S. Tang, "Perspectives in supply chain risk management," *International Journal of Production Economics*, 2006, 103(2), pp. 451-488.