Proceedings of the 10th International Conference on Applied Informatics
Eger, Hungary, January 30–February 1, 2017. pp. 295–302
doi: 10.14794/ICAI.10.2017.295

On the connection of flow graphs and contingency tables

Péter Takács, Zoltán Ernő Csajbók

Faculty of Health, University of Debrecen
{takacs.peter,csajbok.zoltan}@foh.unideb.hu

Abstract

At the millennium, Pawlak introduced the flow graphs in rough set theory, a graphical framework for reasoning about data. This methodological step extended his theory. Pawlak pointed out that between some flow graph theory concepts such as strength, certainty, coverage factors and some theorems of probability theory, namely, total probability theorem and Bayes' theorem have formal relationships. This paper shows that all these flow graph concepts also can formally be associated with statistical concepts, the contingency tables as well.

Keywords: rough set theory, flow graphs, contingency tables

MSC: MSC68R10, 05C21

1. Introduction

One of the goals of Pawlak's rough set theory [10][11] is to manage the concept of uncertainty. Its basic concept has been extended in several directions since its origin. Great progress has been made in both theoretical and practical directions. Further examination of decision-making rules led Pawlak to introduce flow graphs in rough set theory (rough set flow graphs, in short, or simply flow graphs) [12][13]. Since then, rough set flow graphs have been further improved. For instance, Pattaraintakorn examined the concept of entropy in flow graph [9]; Chitcharoen and Pattaraintakorn introduces a matrix form of flow graphs [3][4]. The purpose of this article is to show the connection between rough set flow graphs and statistical contingency tables.

The paper is organized as follows. Section 2 briefly summarizes some statistical terms about contingency tables, whereas Section 3 defines the basic theory of rough set flow graphs. Section 4 basically creates a new modified definition. The example in this section gives a detailed insight into the relations between the two theoretical parts. Further research directions will be outlined in the Summary.

2. Contingency tables

The most basic method of statistical analysis of discrete variables (measured on nominal or ordinal scales) is the analysis of frequencies, relative frequencies and percentage distributions. The results are usually displayed in tables and graphs. Joint distribution of two or more variables can be represented in contingency tables. However, contingency tables of three or even more variables may be too complex — so, their using is not common. This paper also deals with contingency tables containing two variables.

Let X and Y be two discrete random variables. The possible values of X and Y are the events $A_1, \ldots, A_n$ and the events $B_1, \ldots, B_m$, respectively. Consider an experiment to X and Y. If the events A_i and B_j occur at the same time, we say that the event A_iB_j occurs. Let the experiment carry out N times, and the occurrence frequency of A_iB_j denote by f_{ij}. Arranging the values f_{ij} in a table form is called a contingency table (Figure 1). In this table, the dotted indices indicate row and column sums, i.e., $f_{i.} = \sum_{k=1}^{m} f_{ik}$, $f_{.j} = \sum_{k=1}^{n} f_{kj}$, and $\sum_{i=1}^{n} f_{i.} = \sum_{j=1}^{m} f_{.j} = N$.

		Y					
		B_1	...	B_j	...	B_m	
X	A_1	f_{11}		f_{1j}		f_{1m}	$f_{1.}$
	...						...
	A_i	f_{i1}		f_{ij}		f_{im}	$f_{i.}$
	...						...
	A_n	f_{n1}		f_{nj}		f_{nm}	$f_{n.}$
		$f_{.1}$	...	$f_{.j}$	...	$f_{.m}$	N

Figure 1: Common notations in contingency tables

Three basic rates may be identified in a contingency table:
(a) the percentages of the row sums: $f_{ij}/f_{i.}$;
(b) the percentages of the column sums: $f_{ij}/f_{.j}$;
(c) the percentage of the total sum: f_{ij}/N.

3. Rough set flow graphs – basic notions and notations

This section discusses the basic concepts of Pawlak's original rough set flow graphs. It follows the terminology and notations of Pawlak [14].

Definition 3.1. A flow graph is a directed, acyclic, finite graph $G = (N, \mathcal{B}, \varphi)$. N is a set of *nodes*. $\mathcal{B} \subseteq N \times N$ is a set of *directed branches*. $\varphi : \mathcal{B} \to \mathbb{R}^{>0}$ is a *flow function* ($\mathbb{R}^{>0}$ is the set of positive real numbers). $\varphi(x, y)$ models a *throughflow* from x to y $(x, y \in N; (x, y) \in \mathcal{B})$.

A node may have input and/or output nodes. For $x \in N$, the sets of nodes $I(x) = \{y \in N : (y, x) \in \mathcal{B}\}$ and $O(x) = \{y \in N : (x, y) \in \mathcal{B}\}$ are the *input* and the *output of the node* $x \in N$, respectively.

Inputs and outputs of nodes can be aggregated to the whole graph. The sets $I(G) = \{x \in N : I(x) = \emptyset\}$ and $O(G) = \{x \in N : O(x) = \emptyset\}$ are the *input* and the *output of the flow graph* G. The union of input and output nodes of G is referred to as the *external nodes* of G, and $N \setminus (I(G) \cup O(G))$ is the *internal nodes* of G.

The concepts inflow and outflow can also be defined for a node and the whole graph as well. *Inflow of a node* $x \in N$ is a function that summarizes the flow values for the input nodes of x: $\varphi_+(x) = \sum_{y \in I(x)} \varphi(y, x)$. *Outflow of a node* $x \in N$ is a function that summarizes the flow values for the output nodes of x: $\varphi_-(x) = \sum_{y \in O(x)} \varphi(x, y)$. *Inflow of the whole graph* G is a function that summarizes the flow values for the input nodes of graph G: $\varphi_+(G) = \sum_{x \in I(G)} \varphi_-(x)$. *Outflow of the whole graph* G is a function that summarizes the flow values for the output nodes of graph G: $\varphi_-(G) = \sum_{x \in O(G)} \varphi_+(x)$.

In the rough set flow graphs, it is assumed that for each internal node x the following equation holds: $\varphi_+(x) = \varphi_-(x) = \varphi(x)$. $\varphi(x)$ is called the *throughflow of a node* x. Similarly, for the whole graph G, we have $\varphi_+(G) = \varphi_-(G) = \varphi(G)$. $\varphi(G)$ is the *throughflow of the graph* G. Reasonably, these equations are called *flow conservation equations*.

Definition 3.2. A *normalized flow graph* is such a directed, acyclic, finite graph $G = (N, \mathcal{B}, \sigma)$, where N is a set of nodes, $\mathcal{B} \subseteq N \times N$ is a set of directed branches, and σ is a *normalized flow function* which is defined as follows:

$$\sigma : \mathcal{B} \to [0, 1], \quad \sigma(x, y) \mapsto \frac{\varphi(x, y)}{\varphi(G)}.$$

The value $\sigma(x, y)$ is called the *strength of the directed branch* $(x, y) \in \mathcal{B}$.

Clearly, $0 \leq \sigma(x, y) \leq 1$.

In normalized flow graphs, normalized inflow and outflow functions can also be defined. *Normalized inflow* and *outflow of a node* $x \in N$ are:

$$\sigma_+(x) = \sum_{y \in I(x)} \sigma(y, x); \qquad \sigma_-(x) = \sum_{y \in O(x)} \sigma(x, y).$$

Normalized inflow and *outflow of the whole normalized flow graph* are:

$$\sigma_+(G) = \sum_{x \in I(G)} \sigma_-(x); \qquad \sigma_-(G) = \sum_{x \in O(G)} \sigma_+(x).$$

Normalized flow conservation equations are also required in normalized flow graphs. For any internal node x, $\sigma_+(x) = \sigma_-(x) = \sigma(x)$ holds. $\sigma(x)$ is the *normalized throughflow of the node* x. For the whole graph G, $\sigma_+(G) = \sigma_-(G) = \sigma(G)$ holds, where $\sigma(G) = 1$.

The following two notions will be fundamental in the following.

Definition 3.3. Let $G = (N, \mathcal{B}, \sigma)$ be a normalized graph. For any directed branch $(x, y) \in \mathcal{B}$, $cer(x, y) = \dfrac{\sigma(x, y)}{\sigma(x)}$ and $cov(x, y) = \dfrac{\sigma(x, y)}{\sigma(y)}$ are called the *certainty factor* and the *coverage factor* of (x, y), respectively.

4. Connection of flow graphs and contingency tables

Relying on the certainty and coverage factors, Pawlak pointed out in [14] that

> "[...] the information flow in a flow graph is governed by Bayes' formula; however, the formula can be interpreted in an entirely deterministic way without referring to its probabilistic character." ([14], p. 35)

Reviewing many examples of research papers, e.g., [1, 2, 5, 7, 8, 15], one might notice the following two observations:

(a) The nodes locating in a column are never connected in the flow graphs pictures. In such a column, a possible value of an attribute is assigned to each node.

(b) The nodes in a column always connect to the nodes of only one other column.

In graph theory, a graph of such type is called the multipartite or k-partite graph. Let $k \geq 2$ be an integer. A graph $G = (N, \mathcal{B})$ is called *k-partite*, if the nodes can be partitioned into k disjoint sets in such a way that every node has its ends in different equivalent classes. In Pawlak's flow graph context, another additional constraint is added to this definition. Namely, every node from a class has to end in the same other class.

Definition 4.1. The *modified Pawlak's flow graph* is a k-partite, directed, acyclic, finite graph $G = (N, k, \mathcal{B}, \varphi)$. $N = T_1 \cup T_2 \cup \cdots \cup T_k$ is a set of nodes, where $T_1, T_2, \ldots, T_k$ are disjoint node sets, and $k \geq 2$ is an integer. $\mathcal{B} \subseteq N \times N$ is a set of directed branches with the following restrictions: (a) For any $i \leq k$, if $x, y \in T_i$, $(x, y) \notin \mathcal{B}$. (b) For any $i \leq k$, there is the only $i \neq j \leq k$ in such a way that $(x, y) \in \mathcal{B}$ for all $x \in T_i$ and some nodes $y \in T_j$. Such T_i, T_j pairs are called *related*. $\varphi : \mathcal{B} \to \mathbb{R}^{>0}$ is the flow function.

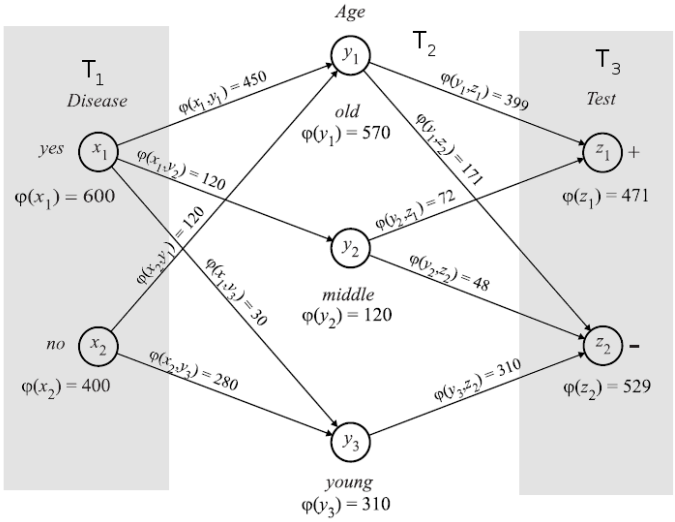
Similar modifications can also be made for normalized flow graphs.

The above definition gives an opportunity to draw a parallel between the concepts of flow graphs and contingency tables.

Let T_p, T_q $(p, q \leq k \in \mathbb{N},\ p \neq q)$ be two related disjoint sets of nodes in the modified flow graph G. It is assumed that $T_p = \{v_1^p, \ldots, v_n^p\} \subset N$ and $T_q = \{v_1^q, \ldots, v_m^q\} \subset N$. Let assign an event $A_i \in X$ to the node $v_i^p \in T_p$ $(i = 1, 2 \ldots, n)$, and an event $B_j \in Y$ to the node $v_j^q \in T_q$ $(j = 1, 2, \ldots, m)$. Then it may be drawn a parallel between the throughflow $\varphi(v_i^p, v_j^q)$ from v_i^p to v_j^q and the occurrence frequency f_{ij} of $A_i B_j$. Between other elements of the the two theories may be drawn a parallel in the same way. These are summarized in Table 1, and the Example 4.2 gives an illustrative example.

Contingency Table	Flow Graph G
f_{ij}, $f_{i.}$, $f_{.j}$	$\varphi(v_i^p, v_j^q)$, $\varphi(v_i^p)$, $\varphi(v_j^q)$
$\frac{f_{ij}}{N}$, $\frac{f_{i.}}{N}$, $\frac{f_{.j}}{N}$	$\sigma(v_i^p, v_j^q) = \frac{\varphi(v_i^p, v_j^q)}{\varphi(G)}$, $\sigma(v_i^p) = \frac{\varphi(v_i^p)}{\varphi(G)}$, $\sigma(v_j^q) = \frac{\varphi(v_j^q)}{\varphi(G)}$
$\frac{f_{ij}}{f_{i.}}$, $\frac{f_{ij}}{f_{.j}}$	$cer(v_i^p, v_j^q) = \frac{\sigma(v_i^p, v_j^q)}{\sigma(v_i^p)}$, $cov(v_i^p, v_j^q) = \frac{\sigma(v_i^p, v_j^q)}{\sigma(v_j^q)}$

Table 1: Parallel notions of contingency tables and flow graphs.
$G = (N, k, \mathcal{B}, \varphi)$; $v_i^p \in T_p, v_j^q \in T_q$; $(v_i^p, v_j^q) \in \mathcal{B}$;
$T_p, T_q \subset N$ $(T_p \cap T_q = \emptyset)$

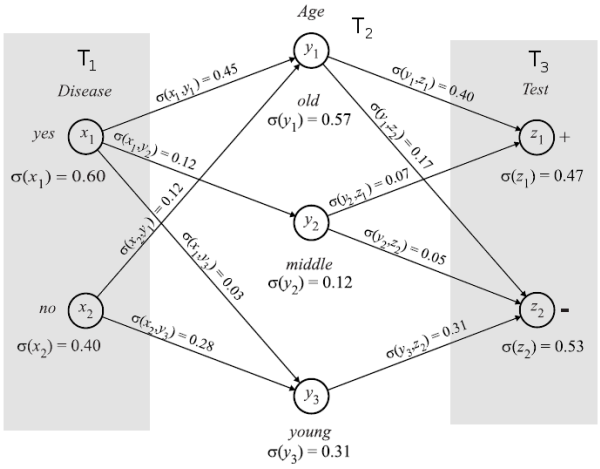


		Y			
		B1	B2	B3	
X	A1	450	120	30	600
	A2	120	0	280	400
		570	120	310	1000

		Z		
		C1	C2	
Y	B1	399	171	570
	B2	72	48	120
	B3	0	310	310
		471	529	1000

Figure 2: Flow graph [14] and the proper contingency tables

Example 4.2. Pawlak's article [14] illustrates the basic concepts of the flow graph theory by an example of a group of 1000 patients. They are examined according to presence of the disease ($T_1 = \{yes, no\}$), age ($T_2 = \{old, middle, young\}$) and test results ($T_3 = \{+, -\}$) – these are the nodes in the flow graph. The grouped nodes are shown in Figure 2. $\varphi(x_1)$ indicates that 600 people are affected by the disease and $\varphi(x_2) = 400$ are not. This can be seen in the sum of the first and second rows in the corresponding contingency table - these are the occurrence frequency of events A_1 and A_2. The distribution of age and test results are given by



		Y			
		B_1	B_2	B_3	
X	A_1	450/1000=0.45	120/1000=0.12	30/1000=0.03	600/1000=0.6
	A_2	120/1000=0.12	0/1000=0	280/1000=0.28	400/1000=0.4
		570/1000=0.57	120/1000=0.12	310/1000=0.31	1000/1000=1

		Z		
		C_1	C_2	
Y	B_1	399/1000=0.399	171/1000=0.171	570/1000=0.57
	B_2	72/1000=0.072	48/1000=0.048	120/1000=0.12
	B_3	0/1000=0	310/1000=0.31	310/1000=0.31
		471/1000=0.471	529/1000=0.529	1000/1000=1

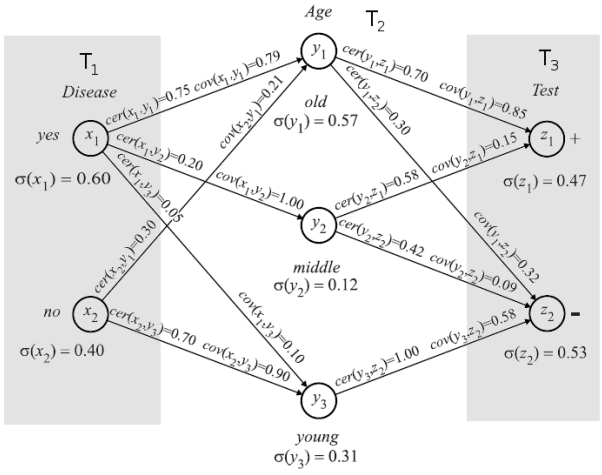
Figure 3: Normalized flow graph [14] and contingency tables with relative frequencies

the $\varphi(y_1), \varphi(y_2), \varphi(y_3)$ and $\varphi(z_1), \varphi(z_2)$ - which values also appear in the contingency tables, occurrence frequency of events B_1, B_2, B_3 and C_1, C_2. The $\varphi(x, y)$ througflow values appear in the contingency table cells. Figure 3 presents the normalized flow graph, the associated σ values and the relative frequencies. The corresponding value pairs differ only by rounding. The certaninty ($cer(x, y)$) and coverage ($cov(x, y)$) factors are shown in Figure 4. The flow graph factors can also be matched with the corresponding cell values in the tables in these cases as well.

5. Summary

One of the main research goal of the rough set theory is the mathematical modelling human decision process. Flow graph theory represents a new chapter in these studying. Examples in many research publications support the possibility of modifying the definition flow graph. This modified flow graph definition allows to examine the similarity between flow graph theory and contingency tables.
Further studies are possible in the following research directions:

- It is necessary to examine the total effect of the modified definition on the theory (Flow Graph- and Roug Set Theory);



		Y		
		B_1	B_2	B_3
X	A_1	450/600=0.75	120/600=0.2	30/600=0.05
	A_2	120/400=0.3	0/400=0	280/400=0.7

		Y		
		B_1	B_2	B_3
X	A_1	450/570=0.789	120/120=1	30/310=0.097
	A_2	120/570=0.211	0/120=0	280/310=0.903

		Z	
		C_1	C_2
Y	B_1	399/570=0.7	171/570=0.3
	B_2	72/120=0.6	48/120=0.4
	B_3	0/310=0	310/310=1

		Z	
		C_1	C_2
Y	B_1	399/471=0.874	171/529=0.323
	B_2	72/471=0.153	48/529=0.091
	B_3	0/471=0	310/529=0.586

Figure 4: Certainty and coverage factors [14] and the corresponding contingency tables with relative frequencies (concerning row sums and column sums)

- It can be studied the effects of swapping variables.
- Pawlak has introduced η dependency (correlation) factor between nodes. There are many types of dependency indicators for contingency tables as well. The relationship of dependency factors in the two research sectors can be studied.

These questions and the further results make it possible to approach the classical statistical concepts and methods from a different perspective and give new ideas to the Pawlak's flow graph research.

References

[1] Barakat, S. I., Eissa, M. M., EL-Henawy, I., Hybrid rough sets and probabilistic flow graph model in coronary artery disease, *Egyptian Comput Sci J, 33 2009.*

[2] Butz, C. J., Yan, W., Extended Pawlak's flow graphs and information theory, *In: Rough Computing: Theories, Technologies and Applications: Theories, Technologies and Applications, Hassanien, A. E. (Ed.), 2007*, 152–161.

[3] Chitcharoen D., Pattaraintakorn P., Novel matrix forms of rough set flow graphs with applications to data integration, *Computers and Mathematics with Applications, Volume 60, Issue 10, November 2010*, 2880–2897.

[4] Chitcharoen, D. Pattaraintakorn, P., Mathematical sub-flow graphs with application to social networks analysis, *2014 International computer science and engineering conference: 2014 International computer science and engineering conference (ICSEC 2014), 30 July-1 August, 2014, Khon Kaen, Thailand*, 261–266.

[5] Ćirović, G., Plamenac, D., The flow networks approach to decision analysis in construction industry, *Yugoslav Journal of Operations Research 17 (2007), Number 1*, 107–124.

[6] Diestel, R., Graph Theory, *Springer-Verlag, Heidelberg Graduate Texts in Mathematics, Volume 173, Electronic Edition 2000: http://diestel-graph-theory.com (2017.05.21.)*

[7] Lui, H., Sun, J., Zheng, H., Lui, L., Extended Pawlak's flow graphs and information theory, *Gavrilova, M.L; et al.(Ed.). Transactional on Computational Science, Springer-Verlag, LNCS, 5540, 2009*, 220–236.

[8] Pal, S. K., Chakraborty, D. B., Granular Flow Graph, Adaptive Rule Generation and Tracking, *IEEE Trans Cybern. 2016 Aug 26.*

[9] Pattaraintakorn, P., Entropy measures of flow graphs with applications to decision trees, *Proceedings of the 4th International Conference on Rough Sets and Knowledge Technology*, LNCS 5589 (2009), 618–625.

[10] Pawlak, Z., Rough sets, *International Journal of Computer and Information Sciences* Vol. 11 (1982), 341–356.

[11] Pawlak, Z., Rough Sets: Theoretical Aspects of Reasoning about Data, *Kluwer Academic Publishers*, Dordrecht (1991).

[12] Pawlak, Z., In Pursuit of Patterns in Data Reasoning from Data - The Rough Set Way, *Rough Sets and Current Trends in Computing: Third International Conference, RSCTC 2002 Malvern, PA, USA, October 14–16, 2002 Proceedings*, 1–9.

[13] Pawlak, Z., Flow graphs - A new paradigm for data mining and knowledge discover, *Proceedings of the JAIST Forum 2004 - Technology Creation Based on Knowledge Science: Theory and Practice, jointly with 5th International Symposium on Knowledge and Systems Sciences (KSS'04) Japan Advanced Institute of Science and Technlogy (JAIST)*, 147–153.

[14] Pawlak, Z., Flow graphs and data mining, *In Transactions on Rough Sets III, James F. Peters and Andrzej Skowron (Eds.), Springer-Verlag, Berlin, Heidelberg, 2005*, 1–36.

[15] Sutoyo, E., Mungad, M., Hamid, S., Herawan, T., An Efficient Soft Set-Based Approach for Conflict Analysis, *PLoS ONE 11(2): e0148837, 2016.*