Pawlak Collaboration Graph of the Second Kind and Its Properties

Zbigniew Suraj, Piotr Grochowalski, and Łukasz Lew

Chair of Computer Science, University of Rzeszów, Poland
{zsuraj,piotrg,lew}@univ.rzeszow.pl

Abstract. In the paper we study some properties of Pawlak collaboration graph of the second kind by means of statistical, graph-theoretical and social network methods. In order to build such graph we use data collected in the Rough Set Database System. Pawlak collaboration graph contains data, among others, on Z. Pawlak, his co-authors and the latter ones' co-authors, etc. In the proposed graph we put an edge between two vertices if the authors have a joint paper, with no other authors. Pawlak collaboration graph can be treated as an example of a large social network. Analyzing this data by means of own computer program, we discover hidden patterns of collaboration represented by this graph which can be interesting for the rough set community and others.

Keywords: large social networks, collaboration graph, cores, lords, rough sets, RSDS system.

1 Introduction

Currently, information gaining particular popularity is the one coming from social networks [4],[16]. In the paper we study some statistical and graph-theoretical properties of Pawlak collaboration graph of the second kind, which is another example of a large social network in comparison to the one discussed in [14],[15]. The main goal of this paper is to define a new model of Pawlak collaboration graph and present the results of analyses of the graph. In [15] the discussion concerning the characteristics of Pawlak collaboration graph has been based on linking two authors (vertices) in the graph if they have written a joint paper, whether or not other authors were involved. In this paper we define Pawlak collaboration graph in such a way that we put an edge between two vertices if the authors have a joint paper, with no other authors. According to this definition, for example, the first co-author of this paper does not have an edge with Pawlak in the graph, since his only joint publication with Pawlak was a three-author paper with Andrzej Skowron and James Peters as well. (But Skowron has still an edge with Pawlak, since some of his joint papers are with Pawlak alone.) Thus, this new definition of Pawlak collaboration graph is more restrictive than the previous one. In order to build such graph we use data collected in the Rough Set Database System [17]. Analyzing our data we discover new hidden patterns of collaboration among members of the rough set community which can be interesting for this community and others. The paper is a follow-up of our earlier

CONCURRENCY, SPECIFICATION AND PROGRAMMING
M. Szczuka et al. (eds.): Proceedings of the international workshop CS&P 2011
September 28-30, Pułtusk, Poland, pp. 512-522

research presented in [11],[13],[14],[15] and related to the problem in question. This paper not only provides a new model of collaboration but also it is enriched with new concepts and techniques from social network analysis (cf. [2]). We assume that the reader is familiar to the basic notions of statistics as well as graph theory (see for example [18],[19]).

The paper is organized as follows. Section 2 provides the definitions of Pawlak number and Pawlak collaboration graph of the second kind. Section 3 presents the results of basic statistical and graph-theoretical analyses of Pawlak collaboration graph of the second kind. Section 4 gives further characteristics of the truncated Pawlak collaboration graph based on cores and lords. Section 5 includes outcome and some considerations on future work.

2 Pawlak Collaboration Graph

In order to reveal a social phenomenon of collaboration in rough set research, we defined the collaboration graph in the paper [11]. In the considered graph the vertices represent all researchers (rough set paper authors [17] in particular), whereas the edges represent collaboration relations between two given authors. Two vertices of the graph are joined by an edge if the two authors have had a joint research paper published, with or without other co-authors. A single edge fixed between two authors in the graph means one or more co-publications. The structure of the collaboration graph together with its basic properties have been presented in [11]. In order to characterize existing collaboration between the rough set community members more precisely we defined a subgraph of the collaboration graph with a distinguished vertex corresponding to Pawlak called Pawlak collaboration graph in [14]. An extended analysis of its basic properties has been presented in [15]. In [14] the discussion concerning the characteristics of Pawlak collaboration graph has been based on linking two authors (vertices) in the graph if they have written a joint paper, whether or not other authors were involved. In this paper we define a new kind of Pawlak collaboration graph in such a way that we put an edge between two vertices if the authors have a joint paper, with no other authors (Therefore we call Pawlak collaboration graph introduced in the paper [14] ”Pawlak collaboration graph of the first kind” and the collaboration graph defined in this paper ”Pawlak collaboration graph of the second kind” in order to make a distinction. Analogously, Pawlak number defined in this paper we call Pawlak number of the second kind, whereas the number defined in [14] we call Pawlak number of the first kind).

Before introducing a definition of Pawlak collaboration graph of the second kind, we need the one of Pawlak number of the second kind. *Pawlak number of the second kind* n_P (in short Pawlak number) of an author is defined as follows: Zdzisław Pawlak himself gets $n_P = 0$; people who have written a joint paper with Zdzisław Pawlak with no other authors, get $n_P = 1$; and their co-authors, with Pawlak number not yet defined, $n_P = 2$; etc.

Pawlak numbers can be interpreted as distances of vertices from Pawlak vertex (number of edges in the shortest path joining a given vertex with Pawlak vertex). In Figure 1 the number of people with Pawlak number 0,1,..., 8 is shown according to the data collected in the RSDS database [17].

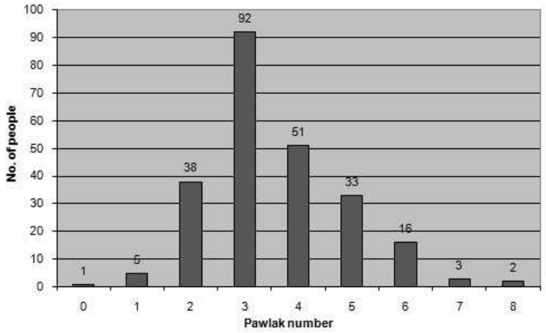

Fig. 1. The distribution of Pawlak numbers

Thus the median of Pawlak number is 3; the mean (the average distance in Pawlak graph between any author and Pawlak) is 3.56, and the standard deviation is 1.31. In our case the standard deviation is low, it indicates that the data points tend to be very close to the mean. This explains that a large number of authors (about 60 percent) obtain Pawlak number within 1.31 of the mean [2.25,4.87] (one standard deviation), and almost all authors (about 98 percent) get Pawlak number within 2.62 of the mean [0.94,6.18] (two standard deviations).

Figure 2 shows Zdzisław Pawlak and the names of all authors with $n_P = 1$.

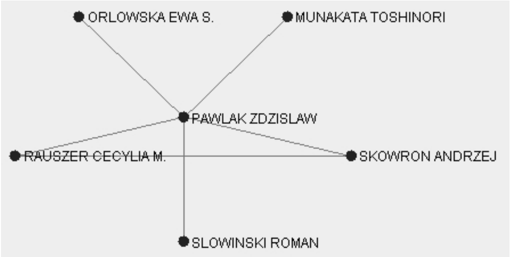

Fig. 2. The names of authors with $n_P = 0, 1$

Remark. In order to distinct the authors with $n_P = 1$ their whole names are written in capital letters.

Now we are ready to define Pawlak collaboration graph of the second kind. A graph $G = (V, E)$, where V is a set of vertices representing authors known in our database RSDS with $n_P \leq 8$, and E is a set of edges connecting two authors, if they wrote a joint paper with no other authors, and at least one of them has $n_P \in \{0, 1, \ldots, 7\}$. The graph G is called *Pawlak collaboration graph of the second kind* (in short *Pawlak graph*). Currently, the data on collaboration among authors with $n_P = 9$ is not available in our database, yet. By removing Zdzisław Pawlak himself and his connections from Pawlak graph G we get the so called *truncated Pawlak graph* G' (shortly *truncated Pawlak graph*).

The data used in our experiments covers the period from 1981 to 2010, inclusive. The last, 2010, edition of the graph G contains 241 vertices and 268 edges (see Table 1), and the graph G' has 240 vertices and 263 edges. There are 538 vertices outside the graph G, which are ignored for the purpose of the analysis; all in all, they do not collaborate with so called Pawlak research group. Other graph-theoretical properties of G' provide insight into the interconnectedness of the rough set researchers. There are 2 connected components in G'. One of them is largest (contains 238 authors) and the other one - extremely small (2 authors). Next, we just concentrate on the largest component of G'. The diameter (maximum distance between two vertices) of the largest component is 12 and the radius (minimum eccentricity of a vertex, with an eccentricity defined as the maximum distance from that vertex to any other vertex) is 6. For any fixed vertex u in the largest component, we can ask about the shape of the distance distribution from u to the other 237 vertices in this component. The distance from u to v is certainly Pawlak number of v, when u is Pawlak. It would be interesting to determine the shape of the distance distribution from a given u to other vertices in the largest component of G', and compare it with the results presented in [6].

Table 1 shows the evolution of Pawlak graph G over time. It is clear that the total vertex number in the graph G varies from 207 (in 2006) to 241 (in 2010). This number is increasing distinctly. A similar tendency can be observed for the edge numbers (see also Figures 3 and 4). It means that the number of rough set authors is growing up as well as the number of joint papers written with no other authors.

Table 1. The evolution of Pawlak graph G over time

	$n_P = 0, 1$		$n_P = 2$		$n_P = 3$		$n_P = 4$		$n_P = 5$		$n_P = 6$		$n_P = 7$		$n_P = 8$		Total	
Year	$\|V_{01}\|$	$\|E_{01}\|$	$\|V_2\|$	$\|E_2\|$	$\|V_3\|$	$\|E_3\|$	$\|V_4\|$	$\|E_4\|$	$\|V_5\|$	$\|E_5\|$	$\|V_6\|$	$\|E_6\|$	$\|V_7\|$	$\|E_7\|$	$\|V_8\|$	$\|E_8\|$	$\|V\|$	$\|E\|$
2006	6	42	36	91	78	34	31	28	25	32	28	2	2	1	1	0	207	230
2007	6	44	38	98	85	37	34	29	26	33	29	3	3	1	1	0	222	245
2008	6	44	38	103	90	56	47	35	31	17	16	3	3	2	2	0	233	260
2009	6	44	38	105	92	60	51	37	33	17	16	3	3	2	2	0	241	268
2010	6	44	38	105	92	60	51	37	33	17	16	3	3	2	2	0	241	268

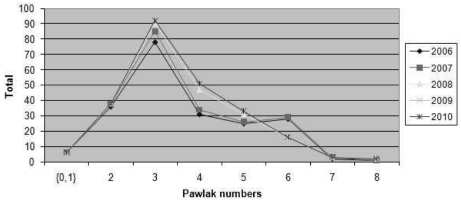

Fig. 3. The evolution of total number of vertices in Pawlak graph G over time

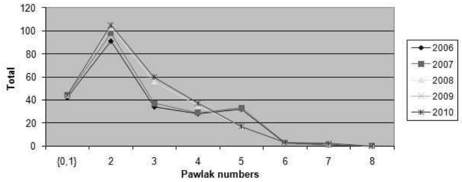

Fig. 4. The evolution of total number of edges in Pawlak graph G over time

3 Basic Analysis of Pawlak Graph

Let us turn now to the issue of collaboration in the rough set research. Firstly, we provide basic statistics of Pawlak graph G, then more advanced graph-theoretical analysis of its properties.

In Table 2 some statistics about the number of co-authors - vertex degrees in Pawlak graph G - are presented. As Table 2 shows, the average degree (number of co-authors for an author) decreases distinctly from 8.33 on level 1 (i.e., Pawlak number is equal to 1) to 1 on level 8. A similar tendency can also be observed for the maximum degrees.

Table 2. Basic statistics on degrees in Pawlak graph G

	$n_P \in \{0,1\}$	$n_P = 1$	$n_P = 2$	$n_P = 3$	$n_P = 4$	$n_P = 5$	$n_P = 6$	$n_P = 7$	$n_P = 8$
Minimum	2	2	1	1	1	1	1	1	1
Median	9.5	8.0	5.5	4.0	4.5	4.0	2.5	2.0	1.0
Average degree	9.0	8.33	4.0	1.76	1.82	1.61	1.19	1.67	1.0
Maximum	22	22	22	23	13	9	4	3	1
Maximizer	(1)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)

Legend: (1) - SKOWRON, A.; (2) - Grzymała-Busse, J.; (3) - Ziarko, W.; (4) - Lin, T.-Y.; (5) - Raś, Z.; (6) - Kryszkiewicz, M.; (7) - Rybiński, H.; (8) - Hong, T. and Strąkowski, T.

According to the number of co-authors the top ten authors are presented in Table 3.

Table 3. Top ten authors according to the number of co-authors

Author	No. of co-authors	Core
1. Ziarko, Wojciech	23	2
2. Grzymała-Busse, Jerzy	22	2
3. SKOWRON, ANDRZEJ	21	2
4. Yao, Yiyu	14	2
5. Lin, Tsau Young	13	2
6. Suraj, Zbigniew	11	2
7. ORŁOWSKA, EWA	10	2
8. Peters, James	10	2
9. Stepaniuk, Jarosław	10	2
10.Raś, Zbigniew	9	1

Remark. The definition of a core is given below.

The two authors with the highest degree in G', Ziarko Wojciech and Grzymała-Busse Jerzy, have written 2 articles together. There are 8 authors with whom both of them are co-authors. Their common co-authors are presented in Table 4.

Table 4. The common co-authors together with the number of common publications

The pairs of co-authors	No. of common publications
Ziarko - Grzymała-Busse	2
Grzymała-Busse - SKOWRON	2
Grzymała-Busse - Yao	1
SKOWRON - Stepaniuk	24
SKOWRON - Suraj	11
SKOWRON - Peters	8
Yao - Lin	2

The distance distribution in G' (number of edges in the shortest path joining the vertices) of other authors taken from Ziarko and Grzymała-Busse is given in Table 5. We can see that Ziarko's co-authors are more collaborative (cf. Average).

4 Collaboration and Influence

In network analysis several measures of collaboration between so called actors can be noted as well as strong influence on their neighborhoods of a given social

Table 5. The distance distribution from Ziarko and Grzymała-Busse

Distance	from Ziarko	from Grzymała-Busse
0	1	1
1	23	22
2	42	62
3	64	91
4	64	32
5	19	19
6	16	11
7	9	0
Sum	238	238
∞	2	2
Average	3.40	2.97

network [2]. We provide experimental results regarding two measures of collaboration and one measure of influence on the rough set community. These measures are defined by using the following notions coming from the graph theory: a k-core introduced by S.B. Seidman [10], a vertex degree, and a lord considered in [2].

4.1 Cores

The structure of large networks can be revealed by partitioning them into smaller units, which are easier to handle. One of such decompositions (apart from connectivity components) is based on so called *cores*.

In a given graph $G = (V, E)$ a subgraph $H_k = (W, E|W)$ induced by the set $W \subseteq V)$ is a *k-core*, or *core of order k*, if and only if $\forall v \in W : \deg_{H_k}(v) \geq k$, and H_k is the maximum subgraph with this property. In other words, vertices belonging to a k-core must be linked to at least k other vertices of the core. The core of maximum order is also called the *main core*. The core number of vertex v, denoted by $core(v)$, is the highest order of a core that contains this vertex.

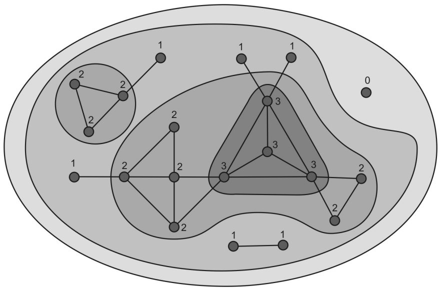

Fig. 5. A core decomposition of a given graph

Figure 5 presents an example of core decomposition of a given graph. From the figure, representing 0, 1, 2 and 3 core, we can see that the cores are nested (i.e., if $i < j$ then $H_j \subseteq H_i$) and they are not necessarily connected subgraphs. It is worth mentioning that there is an efficient algorithm for determining the core hierarchy proposed by [3]. Its idea is based on a simple property:

Property. Deleting recursively all vertices together with edges incident with them, of degree less than k, from a given graph $G = (V, E)$ we obtain the remaining graph being the k-core.

Our experiments prove that the main core in Pawlak graph G consists of 44 vertices. Moreover, its order, similarly to the one of truncated Pawlak graph G', is 2. In Table 6 the distribution of author number in k-cores (second column) and the distributions of co-author number in k-cores for selected members are presented.

Table 6. The distribution of co-author number in cores

Core	No. of authors	Pawlak, Z.	Ziarko, W.	Grzymała-Busse, J.	SKOWRON, A.
2	44	4	5	5	14
1	197	5	23	22	21
Sum	241	9	28	27	35
Average	1.18	1.44	1.18	1.19	1.4

As Table 6 shows, the maximal number of co-authors in 2-core belongs to A. Skowron, and in 1-core to W. Ziarko.

Before introducing a definition of collaboration measures based on the k-core and the vertex degree notions, we need one more notation and two definitions regarding the average degree of all co-authors as well as the average core number of all co-authors.

Let $G = (V, E)$ be a graph, and $v \in V$. By $N(v)$, $\overline{deg}(v)$, $\overline{core}(v)$ we denote, respectively: 1) a set of all neighborhoods of vertex v,

$$N(v) = \{u \in V : (v, u) \in E\}, \tag{1}$$

2) the average degree of vertex v,

$$\overline{deg}(v) = \begin{cases} 0, & N(v) = \emptyset \\ \frac{1}{|N(v)|} \sum_{u \in N(v)} deg(u) & otherwise, \end{cases} \tag{2}$$

where $deg(v)$ is the degree of vertex v; 3) the average core number of vertex v,

$$\overline{core}(v) = \begin{cases} 0, & N(v) = \emptyset \\ \frac{1}{|N(v)|} \sum_{u \in N(v)} core(u) & otherwise, \end{cases} \tag{3}$$

where $core(v)$ is the highest value k for v such that it belongs to a k-core.

In our experiments, the measures of author's collaborativeness are two parameters $coll_1$ and $coll_2$ defined respectively as follows:

$$coll_1(v) = \begin{cases} 0, & deg(v) = 0 \\ \frac{deg(v)}{\overline{deg}(v)} & otherwise \end{cases} \quad and \quad coll_2(v) = \begin{cases} 0, & core(v) = 0 \\ \frac{core(v)}{\overline{core}(v)} & otherwise. \end{cases} \tag{4}$$

The parameter $coll_2$ was introduced in [1]. Let us point out that the high values of $\overline{core}$ and $\overline{deg}$ imply that a 'central' author mainly collaborates with other 'central' authors. Both parameters in question measure the openness of the author v towards 'peripheral' authors. However, the second one is more rigorous than the first. In Table 7 and Table 8 the top ten most collaborative authors w.r.t. $coll_1$ and $coll_2$ are given, respectively. As these tables show, Wojciech Ziarko is the most collaborative author for $coll_1$, and Jarosław Stepaniuk for $coll_2$.

Table 7. The top ten most collaborative authors w.r.t. $coll_1$

Author	deg	$\overline{deg}$	$coll_1$
1. Ziarko, Wojciech	23	2.87	8.014
2. Grzymała-Busse, Jerzy	22	3.95	5.570
3. Lin, Tsau Young	13	2.85	4.561
4. ORŁOWSKA, EWA	11	2.55	4.314
5. SKOWRON, ANDRZEJ	22	5.27	4.175
6. Suraj, Zbigniew	11	3.45	3.188
7. Stepaniuk, Jarosław	10	3.2	3.125
8. Yao, Yiyu	14	4.64	3.017
9. SŁOWIŃSKI, ROMAN	8	3.13	2.556
10.Peters, James	10	4.4	2.272

Table 8. The top ten most collaborative authors w.r.t. $coll_2$

Author	$core$	$\overline{core}$	$coll_2$
1. Stepaniuk, Jarosław	2	1.20	1.667
2. Ziarko, Wojciech	2	1.22	1.639
3. Grzymała-Busse, Jerzy	2	1.23	1.626
4. ORŁOWSKA, EWA	2	1.27	1.575
5. Tsumoto, Shusaku	2	1.29	1.550
6. Lin, Tsau Young	2	1.31	1.527
7. Chakraborty, Mihir	2	1.33	1.504
8. Liu, Qing	2	1.4	1.429
9. Peters, James	2	1.4	1.429
10.Suraj, Zbigniew	2	1.45	1.379

4.2 Lords

Now, let us analyze Pawlak graph G' in order to find all vertices that have 'strong influence' on their neighborhoods. Such vertices are called *lords* [2]. The algorithm for determining lords is quite simple: at the beginning we assign a degree to each vertex with degree as its initial power. The vertices are set increasingly according to their degrees in order to determine the final power distribution, and then the current vertex power is proportionally transmitted to its stronger neighbors.

The top ten lords according to their powers, computed by applying above procedure on G', are given in Table 9. Again, it is worth pointing out that for lords we obtain almost the same set of authors as presented in Table 3. In G', Wojciech Ziarko is the strongest author.

Table 9. Top ten authors according to their powers in the graph G'

	Author	Power
1.	Ziarko, Wojciech	209.583
2.	SKOWRON, ANDRZEJ	193.25
3.	Grzymała-Busse, Jerzy	133.167
4.	Yao, Yiyu	75.167
5.	ORŁOWSKA, EWA	48.5
6.	Lin, Tsau Young	32.667
7.	Raś, Zbigniew	32.667
8.	Peters, James	30.25
9.	Suraj, Zbigniew	27.0
10.	Zhang, Wen-Xiu	26.0

5 Conclusions

We have presented the results of analysis of Pawlak collaboration graph of the second kind in the paper. The concepts of Pawlak numbers considered both in this paper as well as discussed in [14] have some drawbacks. It seems that they are not entirely rational because a person who wrote p ($p > 1$) joint papers with Pawlak is more closely associated with Pawlak than someone who wrote less than p (cf. [1]). Therefore, a lower Pawlak number should be assigned to the first person than to the other one. We would like to investigate the above issue further, develop a suitable Pawlak collaboration graph together with its characteristics using the approach presented in this very paper. Moreover, we are going to present exemplified by Pawlak collaboration graph regarding the number of joint papers additional techniques for analysis of large social networks and visualizations of their parts. In the experiments our own program called

SNetwork is used to make some analyses and get layouts of selected parts of Pawlak collaboration graph.

Acknowledgment. We wish to thank the anonymous referee for critical remarks and useful suggestions to improve the presentation of the paper.

References

1. Barr, M.: Rational Erdös Number. Department of Mathematics and Statistics, McGill University, Montreal, Canada, 1-4 (2001)
2. Batagelj, V., Mrvar, A.: Some Analyses of Erdös Collaboration Graph. Social Networks 22(2), 173-186 (2000)
3. Batagelj, V., Zaversnik, M.: Cores Decomposition of Networks. Recent Trends in Graph Theory, Algebraic Combinatorics, and Graph Algorithms, September 24-27, 2001, Bled, Slovenia (2001)
4. Freeman, L.: The Development of Social Network Analysis. Empirical Press, Vancouver (2006)
5. Grossman, J.W.: The Erdös Number Project (1996) http://www.oakland.edu/grossman/erdoshp.html
6. Grossman, J.W.: Patterns of Collaboration in Mathematical Research. SIAM News 9 (2002)
7. Knuth, D.: The Stanford GraphBase. Addison-Wesley (1993)
8. Pawlak, Z.: Rough Sets. International Journal of Computer and Information Sciences 11, 341-356 (1982)
9. Pawlak, Z.: Rough Sets - Theoretical Aspects of Reasoning About Data. Kluwer (1991)
10. Seidman, S.B.: Network Structure and Minimum Degree. Social Networks 5, 269-287 (1983)
11. Suraj, Z., Grochowalski, P.: Patterns of Collaborations in Rough Set Research. Studies in Fuzziness and Soft Computing, Vol. 224, pp. 79-92. Springer (2008)
12. Suraj, Z., Grochowalski, P.: The Rough Set Database System. Transactions on Rough Sets VIII, LNCS, vol. 5084, pp. 307-331. Springer (2008)
13. Suraj, Z., Grochowalski, P.: Some Comparative Analyses of Data in the RSDS System. LNAI, vol. 6401, pp. 8-15. Springer (2010)
14. Suraj, Z., Grochowalski, P., Lew, Ł.: Pawlak Collaboration Graph and Its Properties. In: Proc. of the 13th Int. Conf. on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, RSFDGrC-2011, Moscow, Russia, June 25-27, 2011, LNCS, vol. 6743, pp. 365-368. Springer (2011)
15. Suraj, Z., Grochowalski, P., Lew, Ł.: Discovering Patterns of Collaboration in Rough Set Research: Statistical and Graph-Theoretical Approach. In: Proc. of the 6th Int. Conf. on Rough Sets and Knowledge Technology, RSKT-2011, Banff, Canada, October 9-12, 2011, LNAI, vol. 6954. Springer (2011)
16. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press (1994)
17. Website of the RSDS system, http://rsds.univ.rzeszow.pl
18. Weiss, N.A: Introductory Statistics. Addison-Wesley (2009)
19. Wilson, R.J., Watkins, J.J.: Graphs, An Introductory Approach. Wiley (1990)