Facebook profiles clustering

Branko Arsić*, Milan Bašić**, Petar Spalević***, Miloš Ilić***, Mladen Veinović****
* Faculty of Science, Kragujevac, Serbia
** Faculty of Sciences and Mathematics, Niš, Serbia
*** Faculty of Technical Sciences, Kosovska Mitrovica, Serbia
**** Department of Informatics and Computing, Singidunum University, Belgrade, Serbia

brankoarsic@kg.ac.rs, basic_milan@yahoo.com, petar.spalevic@pr.ac.rs, milos.ilic@pr.ac.rs, mveinovic@singidunum.ac.rs

***Abstract*— Internet social networks may be an abundant source of opportunities giving space to the "parallel world" which can and, in many ways, does surpass the realty. People share data about almost every aspect of their lives, starting with giving opinions and comments on global problems and events, friends tagging at locations up to the point of multimedia personalized content. Therefore, decentralized mini-campaigns about educational, cultural, political and sports novelties could be conducted. In this paper we have applied clustering algorithm to social network profiles with the aim of obtaining separate groups of people with different opinions about political views and parties. For network case, where some centroids are interconnected, we have implemented edge constraints into classical *k*-means algorithm. This approach enables fast and effective information analysis about the present state of affairs, but also discovers new tendencies in observed political sphere. All profile data, friendships, fanpage likes and statuses with interactions are collected by already developed software for neurolinguistics social network analysis - "Symbols".**

I. INTRODUCTION

In recent years, social media are said to have had an impact on the public discourse and social communication. Social networks, such as Facebook, Twitter and LinkedIn have been becoming very popular during the last few years. People experience various life events, happy or unfortunate life circumstances and all these negative and/or positive impressions are almost immediately shared online, winning inner peace and friends' support or opinion to the others. A great variety of stances is to be found online, independently from the subject of discussion. This permanently enlarges pool of comments on brands, events, educational or health system and could be used as a baseline for research in quality and service improvement [1]. Nonetheless, social network potentials are widely recognized. Many companies, schools, public institutions, political parties, popular individuals and groups have already created online profiles for gathering and analyzing the data [2]. These data are, afterwards, useful in numerous areas such as marketing, public relations, and any type of a thorough research of public opinion [3].

It is certain that, apart from web crawlers that are crucial for forum research, social networks can yield material for sophisticated analyze in the field of marketing and branding [4]. An advantageous approach to grouping people based on their interests comes from the knowledge of their personal data, such as one`s location, birthday, job and education.

In particular, social media are increasingly used in political context [5][6]. Potential voters share their impressions daily in the form of statuses about upcoming events and present state of affairs, their problems, political stances, agreements or disagreements with political activities, plans, and such like daily subjects. In order to meet the citizens' needs, politicians and spin-doctors extract and analyze the information of interest from the available statuses. Twitter is favorite amongst politicians and other known personalities, and thus seems better for collecting and comparing public opinions. Facebook is the most used social network in Serbia, hence we focused our online political study on Facebook. Moreover, Facebook offers the way of entering into direct dialog with citizens and encouraging political discussions, while Twitter streams short flurry of information while the fresh ones rush in continuously. Two more important differences between Facebook and Twitter are: real life friends vs. connecting with strangers and undirected vs. directed edges between profiles. The undirected edges for nodes equality were also the milestone for Facebook selection, too. The unique possibilities of public opinion research through internet, such as real-time data access, knowledge about people's changing preferences and access to their status messages provide prospect for innovation in this field, contrasting to classical offline ways.

In this paper, we present a procedure for finding and analyzing valuable information related to the specific political parties. Our approach is based on Facebook profiles clustering according to their common friends and interests. Clustering techniques can help us to understand relations between profiles and create a global picture of their traits, and eventually conclude how politicians can have impact on them. For this purpose, we adopted well-known clustering algorithm "*k*-means" for dividing social network profiling separate groups, thus providing a room for profiling potential voters. In precise, algorithm *k*-means is adjusted for graph clustering process in order to form several connected components respecting the similarity between nodes. Collecting and filtering is done by already developed software for neurolinguistics social network analysis - "Symbols"[a], which is described in more details, in Section 3. Other approaches are also present and they are focused on analyzing the structure of the social networks and profiles centrality (e.g. see [7, 8, 9, 10]).

[a] http://symbolsresearch.com

The remainder of the paper is structured as follows. Section 2 gives an overview of the literature. Section 3 presents the details of our software "Symbols". Recent surveys of Facebook popularity in Serbia are highlighted in Section 4. Section 5 describes our research methodology. Section 6 extends the standard k-means from vectors to the nodes of graph. The results are presented in Section 7, while Section 8 concludes the study.

II. Related work

Much of real data could be presented as a network (graph). Objects can be presented as nodes, and relations among them as graph's edges. Based on Facebook users' relationships and fanpage likes we have created a network out of Facebook profiles. The problem of data clustering with constraints is now surpassed with graph-based clustering. In this way each element which is be clustered is represented as a node in a graph and the distance between two elements is modeled by a certain weight on the edge thus linking the nodes [11]. The stronger the relation between objects, the higher the weight is (smaller is the distance), and vice-versa. Graph based clustering is a well-studied topic in the literature, and various approaches have been proposed so far.

In paper [12], the graph edit distance and the weighted mean of a pair of graphs were used for cluster graph-based data under an extension of self-organizing maps (SOMs). In order to determine cluster representatives, the authors in [13] conducted the clustering of attributed graphs by means of Function Described Graphs (FDGs). In later approaches the notion of set median graph [14] was presented. It has been used to represent the center of each cluster. However, better presentation of each cluster data is obtained by the generalized median graph concept [14]. Given a set of graphs, the generalized median graph is defined as a graph that has the minimum sum of distances to all graphs in the set. However, median graph approaches are suffering from exponential computational complexity or are restricted to special types of graphs [15]. It would seem that spectral clustering algorithm [16] appears as a much better solution. This method uses the eigenvectors of the adjacency and other graph matrices to find clusters in data sets represented by graphs. k-means clustering algorithm for graphs was introduced [17], bearing in mind the simplicity and speed of algorithms. In this paper we suggested an extension of classical k-means algorithm for Euclidean spaces [18][19], but implemented in the case of graph (see Section 5).

III. "SYMBOLS" data collection

In this section we give a brief overview of Symbols software and its possibilities. As "glue" between our software and Facebook API we developed a Facebook application SSNA (Software for Social Network Analyses). When users start this app, they are asked for the private data access permission. Upon their agreement, the app calls Facebook API on behalf of users after which valid security token for the next two months is obtained. The data encompasses the following network records:

1) The friendship network: ego network includes the SSNA app users (egos) as nodes and friendship relations between them;
2) The communication network:
   (a) Like relations: by clicking a "like" button, Facebook users can value another person's content (posts, photos, videos);
   (b) Comment relations: Facebook users can leave comments on another person's content;
   (c) Post relations: Facebook users can post on the "wall" of another person to leave non-private messages.
3) Affinity network: Attachments to various fanpages and groups implicating support and agreement within their niche.

This software offers graphical presentation of statistical data for selected political parties based on social network statuses and likes, and many more.

IV. Facebook in Serbia

According to the last researches of Ministry of Trade, Tourism and Telecommunications in Republic of Serbia, 93.4% of Internet users aged 16 to 24 have a profile on the social networks (Facebook, Twitter). Our research paper is based on Facebook audience, because most of the world's population are friendly oriented according to this global Internet social network. Facebook Advertisement service presents potential reach of 3,600,000 people from Serbia for the promotion. If we are to believe the self-reported information from people in their Facebook profiles, about 45% of them are women and 55% are men. Information are only available for people aged 18 and older. The largest age group is currently form 18 to 24 with total of 1 440 000 users, followed by the users in the age form 25 to 34. Faculty (College) level educated people participate in about 66%, whilst high school students participate in about 32%. At the same time, percentage for single and married relationship status is 38% to 42%.

V. Methodology

Our research focuses on the political parties' prevalence in the whole of territory of the Republic of Serbia. According to our figures, the total number of grabbed fanpages is 663925 and it corresponds to a total of 78758 profiles. Among these fanpages, 4095 are placed by their creators in the sphere of politics, while 771 pages have more than three likes. Profiles and fanpages are used for graph construction. Profiles represent graph nodes, while fanpages determine a measure for similarity between profiles, i.e. weight of the edges.

Last social research shows that people on the Internet social networks, such as Facebook, mark interactions with small number of friends compared to the total number of friends (about 8%), while the remaining ones are "passive". Members of the mentioned minority have similar interests, common friends, and acquaintances from diverse events. This kind of Internet behavior leads us toward taking into consideration common pages as well as common friends in order to create graph with strong edges. We have taken into consideration the limited number of pages for every political party according to total number of page likes, because a very large number of fanpages can yield misleading results. Bearing this in mind, we selected ten most numerous fanpages of each political party by searching keywords in the title related to their name, abbreviation and leaders. Lets denote this set of fanpages with S. We limited our examination to the four most popular political parties at this moment.

VI. ADOPTED k-MEANS ALGORITHM

The concept of a sample mean is defined as the mean of the observed samples. The sample mean is well-defined for vector spaces only, and we are often forced to present objects by definite discrete structures such as strings, graphs, sets, etc., where sample mean is not always possible to define. The k-means algorithm is a popular clustering method because of its simplicity and speed [20][21]. Algorithm 1 describes k-means for vectors in order to point out changes with our adaptation for graphs.

Algorithm 1: *k-means algorithm for Euclidian space.*

1. Choose initial centroids $Y = \{y_1, \dots, y_k\} \subset \mathrm{X}$, where X is a set of all vectors and $|Y| = k$.
2. **repeat**
3. assign each $x \in \mathrm{X}$ to its closest centroid $y = y(x) = argmin_{y \in Y} \| x - y \|^2$ of a cluster $C(y)$
4. recompute each centroid $y \in Y$ as the mean of all vectors from $C(y)$
5. **until** *some termination criterion is satisfied*;

As previously mentioned, we did not consider only friends connections for graph construction, but also the same interests and common friends in order to make stronger connections among people. We say that two friends are connected if they have more than three fanpages (four and five have been also tested) and more than four common friends; otherwise we disconnect the edge in graph. Through the same interests and acquaintances, created edges represent strong relations between active friends (Fig. 1). In accordance with these rules, we obtained a graph with 428 nodes and 4448 edges (more than three fanpages and four friends in common, Fig. 2). In a spirit of k-means algorithm, for similarity between connected nodes we used the following function:

$$sim = \frac{1}{\alpha \times \sigma(u, v) + \beta \times \phi(u, v)}$$

where $\sigma(u, v)$ presents structural similarity between nodes [22] and $\phi(u, v)$ the number of chosen fanpages in common for profiles u and v and then divided by total number of pages (40 in our case). The smaller the value of similarity function, the closer the nodes are. Parameters α and β can be used to favour one of the parameters. Here, we considered that $\alpha = \beta = 1$. If we obtained a disconnected graph, we would choose two arbitrary nodes from any separated components and make an edge between them with the smallest similarity value, and so on until the connected graph is obtained. For cluster centers determination we used betweenness centrality as an indicator of a node's centrality in a network [23]. We chose this measure because betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes thus matching the nature of a problem. A node with high betweenness centrality has a large influence on the transfer of items through the network. The Algorithm 2 presents an adaptation of Algorithm 1 for graph paradigm.

Algorithm 2: *k-means algorithm for graphs.*

1. Choose initial centroids $Y = \{y_1, \dots, y_k\} \subset \mathrm{nodes(G)}$, where $nodes(G)$ is a set of graph nodes and $|Y| = k$.
2. **repeat**
3. assign each $x \in nodes(G)$ to its closest centroid $y(x) = argmin_{y \in Y} \sum_{e \in shortest\_path} sim(e)$ of a cluster $C(y)$
4. replace each centroid $y \in Y$ with the node which corresponds to the maximal value of betweenness centrality of all nodes from $C(y)$
5. **until** *number of iterations is equal to* t;

The first step in data clustering is determining a number of clusters k. Generally speaking, number of clusters k is determined in advance according to data sample. The problem we have been solving suggests the fixed cluster number with the value 4. First step is to randomly choose four nodes. In every loop step, an association of all nodes to the nearest centroid is performed. The nearest centroid is determined as a minimal sum of weights along the shortest path between a node and centroids. The next step includes betweenness centrality calculation for every current cluster and the replacing centroids according to the largest coordinate. Calculating the betweenness centrality of all the vertices in a graph is very complex. It is precisely $\Theta(|nodes(G)|^3)$ time-consuming, because it involves calculation of the shortest paths between all pairs of vertices in a graph. We have noticed in numerous experiments that after a few iterations centroids remain the same. This feature has a good influence on algorithm

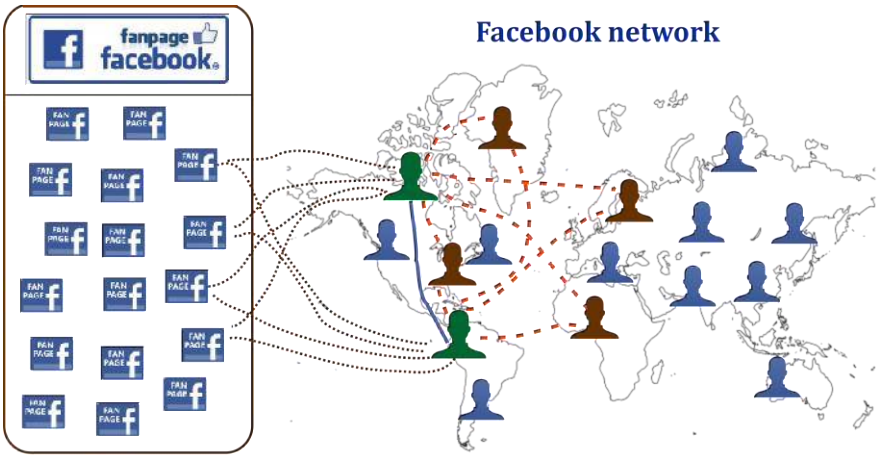

Figure 1. Friends (green) with four fanpages and four friends in common.

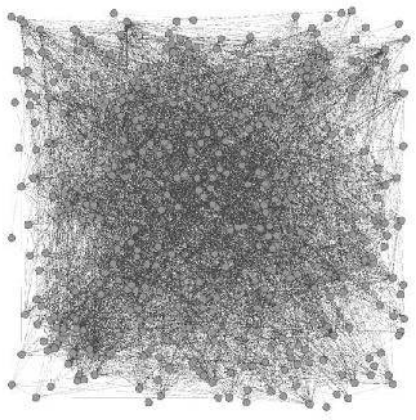

Figure 2. Facebook profiles network, 428 nodes and 4448 edges.

complexity, because we do not need to execute a large number of iterations. Experimental results suggest us to set the number of iterations t from two to four. The calculation of shortest paths between graph nodes in the third step of Algorithm 2 are used for betweenness centrality calculations in the fourth step. This is also relaxation for algorithm calculation complexity. In the following section we give an overview of the experimental results.

VII. Results and discussion

This section is dedicated to experimental results obtained by applying Algorithm 2 on the data collected. Our experiments on profiles are divided into three groups according to the number of fanpages in common: with more than three, four and five fanpages in common. Firstly, we fixed number of clusters to $k = 4$ (number of the most popular political parties in Serbia). Secondly, after the algorithm for clustering is performed in graph constructed of Facebook profiles, for each cluster we have listed all fanpages from S liked by its profiles. Simultaneously, with respect to the cluster, we calculated number of likes for each fanpage listed. A list sample is presented in Table 1.

Based on this list, we determine which political party each cluster represents. Sometimes, it happens that cluster consists of inadequate fanpages, the ones which do not belong to an expected party. If so, the problem of noise is solved with the percentage of contribution calculation for the most dominant fanpages belonging to a political party. If this figure is higher than 80% we relate a cluster with the corresponding party. On the contrary, we mark

TABLE I.
FANPAGES WHICH BELONG TO PROFILES FROM ONE CLUSTER

Fanpage name	Number of likes	Political party
Fanpage 1	2	Party 1
Fanpage 2	2	Party 1
Fanpage 3	2	Party 1
Fanpage 4	2	Party 1
Fanpage 5	2	Party 1
Fanpage 6	2	Party 2

cluster as ”mixed” if the ratio is less than 80% (see Table 1). In almost all cases we had one ”mixed” and three ”clean” clusters. Tables 2, 3 and 4 show the results of experiments for five algorithm starts per group, the percentage of contribution and the number of nodes in the cluster.

TABLE II.
NUMBER OF THE FANPAGES IN COMMON IS GREATER THAN 3. THE NUMBERS OF NODES AND EDGES ARE 428 AND 4448, RESPECTIVELY

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1.	MIXED - 278	98,84% - 89	82,30% - 19	92,68% - 42
2.	MIXED - 375	88,95% - 17	100,0% - 6	95.13% - 30
3.	MIXED - 320	92,54% - 47	86,27% - 17	85,25% - 44
4.	MIXED - 335	97,46% - 12	92,41% - 43	95,56% - 38
5.	MIXED - 325	92,37% - 47	97,43% - 12	84,47% - 44

TABLE III.
NUMBER OF THE FANPAGES IN COMMON IS GREATER THAN 4. THE NUMBERS OF NODES AND EDGES ARE 213 AND 1141, RESPECTIVELY

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1.	MIXED - 142	98,68% - 32	92,42% - 30	82,30% - 9
2.	MIXED - 187	100% - 17	100% - 2	95.93% - 20
3.	MIXED - 167	88,04% - 16	95,93% - 20	92,30% - 10
4.	MIXED – 150	95,93% - 20	96,55% - 40	100% - 3
5.	MIXED - 180	87,32% - 12	95,40% - 18	100% - 3

TABLE IV.
NUMBER OF THE FANPAGES IN COMMON IS GREATER THAN 5. THE NUMBERS OF NODES AND EDGES ARE 93 AND 298, RESPECTIVELY

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1.	MIXED – 68	66,67% - 2	100% - 10	95,18% - 13
2.	MIXED – 75	92,30% - 2	90,67% - 13	58,33% - 3
3.	MIXED - 71	88,89% - 6	100% - 3	95,18% - 13
4.	MIXED – 65	95,18% - 13	100% - 2	99,12% - 13
5.	MIXED - 74	91,30% - 12	100% - 2	96,15% - 5

The largest clusters, consisting of profiles affiliated with different political parties at the same time were indecisive ones. This anomaly can be explained as a consequence of numerous coalitions, both local and global. In this cluster, we noticed that the fanpages of two specific political parties cover the largest part of all fanpages listed population. The two of them dominate alternately, but at all times one political party fanpages contribute between 45% and 60% of the fanpages set, depending on the contents of other corresponding clusters. Even though these results are consistent with the results of online polls conducted on -

"Tvoj stav"[b], and may contain valuable information useful for additional comments, we shall avoid drawing generalized conclusions and will not deal with such clusters. Finally, with these clusters we are able to make a voter's profile for a political party in a simple way.

VIII. CONCLUSION

People share contents about almost every aspect of their life, from opinions on global problems, comments on events, to criticism of political parties and their leaders. These daily online activities encourage the opinion exchange, thus creating political clusters aimed at inspiring certain political actions and coaxing new voters. The goal of this research was to study network ties between profiles according to their common interests. In this paper, we presented a novel graph-based clustering approach which relies on classical k-means algorithm. The algorithm was tested on real Facebook data, and we showed that similar conclusions could be obtained in a faster way when compared to the research conducted by marketing agencies engaged for the same purpose and tasks. We determined three clear clusters for chosen political parties, so that we could distinguish them. The fourth cluster (mixed) consists of about 50% of all the profiles, and this problem remains unsolved. In the future, our efforts would be oriented to its splitting, because undecided group of voters seems to hide important information. The algorithm k-means++ should be a good start [24]. With small modification the same algorithm could be tested on Twitter data. An application upgrade for Twitter profiles will also be our tendency for the future research.

ACKNOWLEDGMENT

This paper was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (scientific projects OI174033, III44006, ON174013 and TR35026).

REFERENCES

[1] C. C. Aggarwal, "An introduction to social network data analytics," in *Social Network Data Analytics*, C. C. Aggarwal, Eds. Springer US, 2011, pp. 1–15.

[2] S. A. Catanese, P. De Meo, E. Ferrara, G. Fiumara and A. Provetti, "Crawling facebook for social network analysis purposes," In *Proceedings of the international conference on web intelligence, mining and semantics*, ACM, pp. 52, 2011.

[3] M. Burke, R. Kraut and C. Marlow, "Social capital on Facebook: Differentiating uses and users," In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp. 571–580, 2011.

[4] B. Arsić, P. Spalević, L. Bojić and A. Crnišanin, "Social Networks in Logistics System Decision-Making," In *Proceedings ot the 2nd Logistics International Conference*, pp. 166–171, 2015.

[5] D. Zeng, H. Chen, R. Lusch and S. H. Li, "Social media analytics and intelligence," *Intelligent Systems*, IEEE, vol. 25, no. 6, pp. 13–16, 2010.

[6] S. Wattal, D. Schuff, M. Mandviwalla and C. B. Williams, "Web 2.0 and politics: the 2008 US presidential election and an e-politics research agenda," *Mis Quarterly*, vol. 34, pp. 669–688, 2010.

[7] S. Catanese, P. De Meo, E. Ferrara, G. Fiumara and A. Provetti, "Extraction and analysis of facebook friendship relations," In *Computational Social Networks*, A. Abraham, Eds. Springer London, 2012, pp. 291–324.

[8] M. G. Everett and S. P. Borgatti, "The centrality of groups and classes," *The Journal of mathematical sociology*, vol. 23, pp. 181–201, 1999.

[9] J. Scott, *Social network analysis*. Sage, London, 2012.

[10] J. Sun and J. Tang, "A survey of models and algorithms for social influence analysis," in *Social Network Data Analytics*, C. C. Aggarwal, Eds. Springer US, 2011, pp. 177–214.

[11] A. K. Jain, M. N. Murty and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[12] S. Günter and H. Bunke, "Self-organizing map for clustering in the graph domain," *Pattern Recognition Letters*, vol. 23, pp. 405–417, 2002.

[13] F. Serratosa, R. Alquézar and A. Sanfeliu, "Synthesis of function-described graphs and clustering of attributed graphs," *International journal of pattern recognition and artificial intelligence*, vol. 16, pp. 621–655, 2002.

[14] X. Jiang, A. Münger and H. Bunke, "An median graphs: properties, algorithms, and applications," *Pattern Analysis and Machine Intelligence*, IEEE Transactions on, vol. 23, pp. 1144–1151, 2001.

[15] H. Bunke, A. Münger and X. Jiang, "Combinatorial search versus genetic algorithms: A case study based on the generalized median graph problem," *Pattern recognition letters*, vol. 20, pp. 1271–1277, 1999.

[16] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, pp. 395–416, 2007.

[17] A. Schenker, *Graph-theoretic techniques for web content mining*, World Scientific, 2005.

[18] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *Pattern Analysis and Machine Intelligence*, IEEE Transactions on, vol. 24, pp. 881–892, 2002.

[19] P. Berkhin, "A survey of clustering data mining techniques," In *Grouping multidimensional data*, J. Kogan, N. Charles and T. Marc, Eds. Springer Berlin Heidelberg, 2006, pp. 25–71.

[20] D. Arthur and S. Vassilvitskii, "How Slow is the k-means Method?" In *Proceedings of the twenty-second annual symposium on Computational geometry*, ACM, pp. 144–153, 2006.

[21] S. Har-Peled and B. Sadri, "How fast is the k-means method?," *Algorithmica*, vol. 41, pp. 185–202, 2005.

[22] X. Xu, N. Yuruk, Z. Feng and T. A. Schweiger, "Scan: a structural clustering algorithm for networks," In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 824–833, 2007.

[23] L. C. Freeman, "A set of measures of centrality based upon betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[24] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics*, pp. 1027–1035, 2007.

[b] http://www.tvojstav.com/page/analysis#analize_mdr