

# Identifying network communities in Bollywood Tweet dataset

Sandeep Ranjan<sup>1</sup>, Dr. Sumesh Sood<sup>2</sup>

<sup>1</sup>PhD Scholar, I.K. Gujral Punjab Technical University, Kapurthala (India)

<sup>2</sup>Assistant Professor, I.K. Gujral Punjab Technical University, Kapurthala (India)

## ABSTRACT

Social Networks are composed of dense communities which can be viewed as symbolic building blocks of the network. Treating social networks as graphs, different parameters can be calculated to identify communities that are distinct from other connected components of the graph. In this research, Tweets for Twitter handles of Bollywood movies released between 1<sup>st</sup> June 2017 and 28<sup>th</sup> July 2017 were fetched to identify communities in the network. This helps to understand the density of the graph. Communities are connected components of the graph where an edge exists between any two nodes of the subgraph. Information spread and the role of communities can be determined by analyzing various graph metrics. Number of connected components in a network symbolizes active disseminators of information.

**Keywords:** - communities, connected components, graphs, opinion mining, social networks

## I. INTRODUCTION

Networks are the best way to represent social, information, technological and biological systems [1]. The nodes of network represent entities and the edges represent relationships amongst various entities. Nodes organize themselves into densely connected communities called network communities or clusters [2]. Networks can be found everywhere, for example a country is divided into provinces, districts, towns and villages. Railway station networks [3], road networks [4], webpage networks [5], biological networks [6], research citation networks [7] and animal society networks [8] are complex networks formed of densely connected nodes. Recently a lot of work has been carried out on social networks, which are formed as a result of the interaction of individuals on an online platform [9,10].

Studies have shown that most networks share a number of distinctive properties. One of these properties is the small world property which states that the average distance between the nodes of a network is short [11]. It means that even in a large network like network of web pages, the average distance between any two nodes is not of a high order. This means that nodes have edges amongst themselves and this gives rise to communities where constituent nodes have shared common properties.

Community detection aims at identifying the groups of nodes that are better connected to each other in contrast to other subgroups of the graph [12]. Social networks are composed of dense communities representing inclination of the general public towards specific topics and actors. Clustering web clients having similar interests and geographically based near to each other can help improve the performance of the World Wide Web. In case of online retail networks like Ebay and Amazon, identifying clusters of clients with similar

interests of purchases, better recommendation models can be developed which can better guide the customers and benefit the retailer by enhancing the business opportunities. In the research presented, dataset was created for Bollywood movies released over a period of two months and graph metrics were calculated to find the individual communities denoted by connected components.

## **II. RELATED WORK**

Analysis of networks has been carried out in a variety of scenarios to provide an alternate approach to problem solving. Zachary's network of the members of the karate club is a well-known example of community detection [13]. The edges represent the outside club activities and interactions of the members, using the community detection algorithms, splitting of the club into two groups can be predicted. Researchers have used community detection in networks for studying animal behavior as social networks are not adversely affected by sampling errors and their effect is also local [14].

Five online social networks Facebook, Twitter, Google+, Flickr and Wikipedia were evaluated using Communities from Edge Structures and Node Attribute (CESNA) to statistically model the network structure-node attributes interactions [15]. Community detection and information spread have been studied in Twitter for the role in the spread of hoax in Indonesia [16]. The study focused on the role of social media in spread of false news, lies and gossip that lead to rumors of death of a famous person in Indonesia. Viral epidemics and large scale opinion sharing leads to success or failure of certain events like TV shows, movie release and music concerts [17].

In the present research, analysis has been carried out on the detection of movie tweet communities. Whenever a Twitter user comments or responds to any movie Twitter handle, edges are created amongst vertices and leads to the evolution of communities. The communities composed of professional critics, viewers and amateur communities affect movie related concepts like box office revenue, movie ratings and help form an image of the movie those reading the content to help them decide whether to watch the movie or not [18]. Twitter based movie rating has been found to be superior over public rating systems like MovieLens and Netflix [19]. The Twitter dataset was found to be more realistic and natural giving nearly real time viewer opinion about newly released movies.

## **III. DATASET CREATION AND OVERVIEW**

NodeXL was used to fetch Tweet datasets for Bollywood Hindi movies released from 1<sup>st</sup> June 2017 to 28<sup>th</sup> July 2017 on a daily basis. Individual movie tweets were fetched in separate MS Excel sheets date wise and a summary sheet was created to save the 7 day data. To avoid redundancy, the filter was applied to Tweet column to get distinct tweets. NodeXL was used to compute the below mentioned parameters for all the movie tweets dataset files.

**Table 1. NodeXL overall metrics worksheet**

| Type of graph                             | Directed/ undirected                                                                                  |
|-------------------------------------------|-------------------------------------------------------------------------------------------------------|
| Vertices                                  | Total vertices in the graph                                                                           |
| Reciprocated Vertex Pair Ratio            | Ratio of number of vertex pairs having bidirectional edges to the number of connected vertex pairs.   |
| Total Edges                               | Total edges in the graph                                                                              |
| Edges containing Duplicates               | Number of edges having any duplicates.                                                                |
| Unique Edges                              | Number of unique edges.                                                                               |
| Self Loops                                | Number of edges forming a loop to vertex.                                                             |
| Single-Vertex Connected Components        | Number of connected components that have only one vertex.                                             |
| Reciprocated Edge Ratio                   | Number of edges that are reciprocated divided by the total number of edges.                           |
| Connected Components                      | Number of connected components in the graph                                                           |
| Diameter                                  | Distance among all vertex pairs, when geodesic distance is the shortest path.                         |
| Maximum Vertices in a Connected Component | Number of vertices in the connected component that has the most vertices.                             |
| Maximum Edges in a Connected Component    | Number of edges in the connected component having the most edges.                                     |
| Modularity                                | A measure of distinguishing denser and sparse connections                                             |
| Graph Density                             | Number of edges divided by the number of graph edges if all the vertices are connected to each other. |

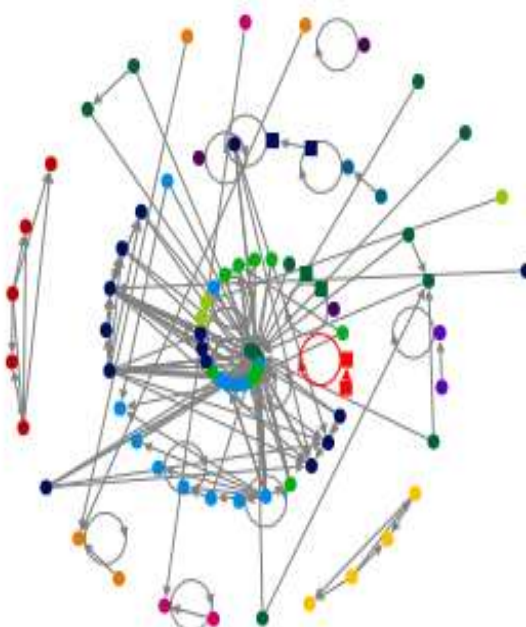
Table 1 describes different network parameters computed by NodeXL and their description. These parameters help to understand network and its structure.

Table 2 shows “Aksar 2” movie tweet sheet for 22<sup>nd</sup> July, 2017 dataset. All the above mentioned fields have been populated with values computed by NodeXL’s graph metrics option.

**Table 2. NodeXL overall metrics worksheet for “Aksar 2” movie tweet data for 22<sup>nd</sup> Nov**

| Metric                            | Value    |
|-----------------------------------|----------|
| Directed/ Undirected              | Directed |
| Number of Vertices                | 83       |
| Number of Unique Edges            | 135      |
| Number of Edges (With Duplicates) | 60       |
| Total Number of Edges             | 195      |
| Number of Self-Loops              | 24       |

|                                                 |             |
|-------------------------------------------------|-------------|
| Vertex Pair Ratio (Reciprocated)                | 0.022222222 |
| Edge Ratio (Reciprocated)                       | 0.043478261 |
| Number of Connected Components                  | 14          |
| Number of Connected Components of Single Vertex | 3           |
| Connected Component Maximum Vertices            | 51          |
| Connected Component Maximum Edges               | 160         |
| Diameter (Geodesic Distance)                    | 4           |
| Geodesic Distance (Average)                     | 2.392738    |
| Density                                         | 0.020276227 |



**Fig.1 Connected components of “Aksar 2” movie tweet data for 22<sup>nd</sup> Nov**

Figure 1 describes the spiral graph of the Aksar 2 movie dataset for 22<sup>nd</sup> November 2017. There is a total of 15 (14+1) connected components in this graph each being represented by a distinct color for differentiation. If any of the vertices is clicked, its corresponding connected component in the graph gets highlighted and the vertex shape is also highlighted. In figure 1, the connected component G14 is selected in the worksheet and thus a red colored solid square appears in the graph.

## IV. EXPERIMENT FRAMEWORK

The dataset has been created by fetching tweets for Twitter handles of Bollywood Hindi movies released from 1<sup>st</sup>

June 2017 to 28<sup>th</sup> July 2017. The tweets for each movie were fetched starting from their release date that normally falls on a Friday to Thursday of next week thus creating dataset of 7 day tweets for each movie released during the said period.

| S. No | Movie Name                    | Twitter handle        | Date (Released on) | Tweet Count |       |       |       |       |       |       | Total | Total Distinct |
|-------|-------------------------------|-----------------------|--------------------|-------------|-------|-------|-------|-------|-------|-------|-------|----------------|
|       |                               |                       |                    | Day-1       | Day-2 | Day-3 | Day-4 | Day-5 | Day-6 | Day-7 |       |                |
| 1     | Sweetie Weds NRI              | sweetiewedsnri        | 02/06/17           | 431         | 457   | 474   | 367   | 295   | 272   | 264   | 2560  | 138            |
| 2     | Mirror Game Film              | mirrorgamefilm        | 02/06/17           | 4905        | 7061  | 7068  | 5030  | 5364  | 6351  | 3641  | 39420 | 678            |
| 3     | Hanuman da Damdaaar           | hanumandadamdaaar     | 02/06/17           | 3855        | 3700  | 3446  | 3485  | 3464  | 2372  | 3383  | 23705 | 1269           |
| 4     | Flat 211                      | flat211               | 02/06/17           | 921         | 991   | 0     | 0     | 902   | 868   | 0     | 3682  | 119            |
| 5     | Dobaara                       | dobaara               | 02/06/17           | 881         | 636   | 4409  | 1090  | 1241  | 1501  | 1634  | 11392 | 2453           |
| 6     | Dear Maya                     | dearmaya              | 02/06/17           | 447         | 4172  | 4396  | 0     | 4363  | 3057  | 4173  | 20608 | 1312           |
| 7     | Behen Hogi Teri               | behenhogiteri         | 09/06/17           | 160         | 3408  | 0     | 3330  | 3453  | 3669  | 3864  | 17884 | 1989           |
| 8     | Raabta                        | raabta                | 09/06/17           | 3129        | 1591  | 0     | 3154  | 911   | 3052  | 3019  | 14856 | 4391           |
| 9     | Love You Family               | loveyoufamily         | 09/06/17           | 42          | 47    | 0     | 0     | 47    | 45    | 46    | 227   | 24             |
| 10    | Bank Chor                     | bankchor              | 16/06/17           | 2562        | 79    | 590   | 754   | 1799  | 1468  | 1466  | 8718  | 1747           |
| 11    | Tubelight                     | tubelight             | 23/06/17           | 3204        | 2050  | 0     | 3166  | 2146  | 4214  | 0     | 14780 | 4657           |
| 12    | Ek Haseena Thi Ek Deewana Tha | ehtedt                | 30/06/17           | 1206        | 7166  | 1497  | 0     | 737   | 409   | 0     | 11015 | 932            |
| 13    | Mom                           | mommovie              | 07/07/17           | 130         | 0     | 212   | 219   | 214   | 216   | 209   | 1200  | 59             |
| 14    | Guest In London               | guestinlondon         | 07/07/17           | 0           | 587   | 0     | 971   | 973   | 981   | 1145  | 4657  | 383            |
| 15    | Jagga Jasoos                  | jaggajasoos           | 14/07/17           | 5281        | 450   | 0     | 3355  | 2435  | 1529  | 731   | 13781 | 5105           |
| 16    | Shab                          | shabthefilm           | 14/07/17           | 4529        | 420   | 0     | 942   | 325   | 513   | 630   | 7359  | 970            |
| 17    | Lipstick Under My Burkha      | lipstickundermyburkha | 21/07/17           | 0           | 360   | 4714  | 5833  | 7756  | 8881  | 2535  | 30079 | 3360           |
| 18    | Munna Michael                 | munnamichael          | 21/07/17           | 0           | 692   | 1396  | 782   | 17467 | 17127 | 17030 | 54494 | 3720           |
| 19    | Raag Desh                     | raagdesh              | 28/07/17           | 465         | 2678  | 3807  | 1529  | 1688  | 2914  | 3572  | 16653 | 1362           |
| 20    | Indu Sarkar                   | indusarkar            | 28/07/17           | 1865        | 8462  | 1792  | 7299  | 8749  | 10629 | 4867  | 43663 | 3123           |
| 21    | Mubarakan                     | mubarakan             | 28/07/17           | 3067        | 1435  | 5122  | 761   | 5337  | 5409  | 4630  | 25761 | 2463           |
| 22    | Baat Company                  | baatcompany           | 28/07/17           | 107         | 141   | 197   | 222   | 224   | 228   | 218   | 1337  | 54             |

Table 3. Movie dataset description

The entries in the table where “0” is entered for a given day Tweet number means there were no Tweets fetched, it is more likely in the cases of movies which are not too popular among social network users. Analyzing movie

tweet datasets as mathematical graphs, different parameters can be calculated [20]. Table 4 contains different network metrics for the “Aksar 2” movie tweets fetched on 22<sup>nd</sup> Nov.

**Table 4. Visual properties of connected components of “Aksar 2” movie tweet data for 22<sup>nd</sup> Nov**

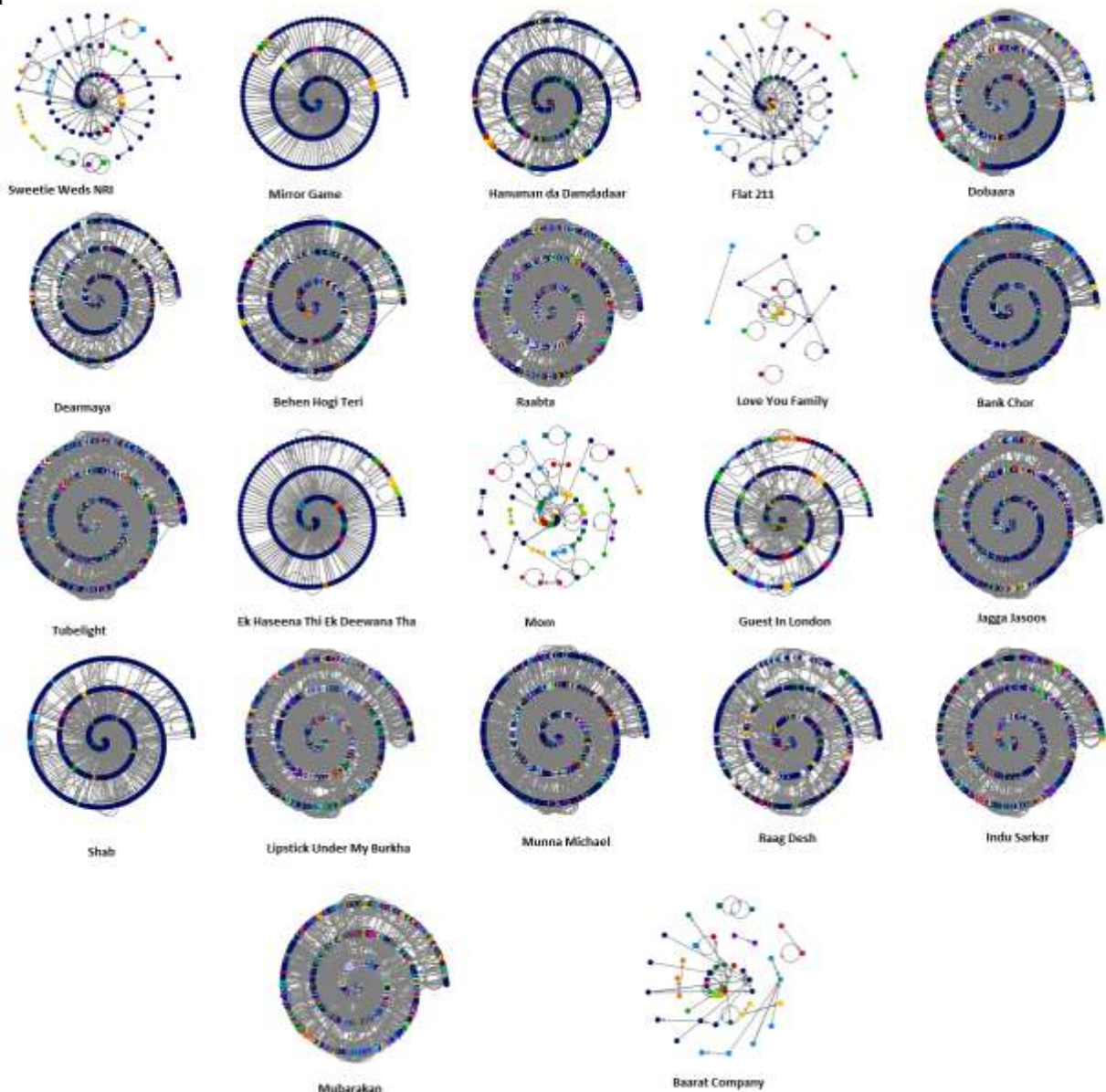
**Table 5. Number of connected components for all the movie tweet datasets**

| Group | Vertex Color | Vertex Shape | ID | Vertices | Unique Edges | Edges With Duplicates | Total Edges | Self-Loops | Reciprocated Vertex Pair Ratio | Reciprocated Edge Ratio | Connected Components | Singl e-Verte x Connecte d Components | Maximum Vertices in a Connected Component |
|-------|--------------|--------------|----|----------|--------------|-----------------------|-------------|------------|--------------------------------|-------------------------|----------------------|---------------------------------------|-------------------------------------------|
| G1    | 0, 12, 96    | Disk         | 3  | 14       | 23           | 7                     | 30          | 2          | 0.087                          | 0.160                   | 1                    | 0                                     | 14                                        |
| G2    | 0, 136, 227  | Disk         | 4  | 14       | 19           | 24                    | 43          | 11         | 0.000                          | 0.000                   | 1                    | 0                                     | 14                                        |
| G3    | 0, 100, 50   | Disk         | 5  | 13       | 17           | 0                     | 17          | 0          | 0.000                          | 0.000                   | 1                    | 0                                     | 13                                        |
| G4    | 0, 176, 22   | Disk         | 6  | 10       | 16           | 3                     | 19          | 0          | 0.000                          | 0.000                   | 1                    | 0                                     | 10                                        |
| G5    | 191, 0, 0    | Disk         | 7  | 5        | 7            | 0                     | 7           | 0          | 0.000                          | 0.000                   | 1                    | 0                                     | 5                                         |
| G6    | 230, 120, 0  | Disk         | 8  | 4        | 4            | 0                     | 4           | 1          | 0.000                          | 0.000                   | 1                    | 0                                     | 4                                         |
| G7    | 255, 191, 0  | Disk         | 9  | 4        | 5            | 0                     | 5           | 0          | 0.000                          | 0.000                   | 1                    | 0                                     | 4                                         |
| G8    | 150, 200, 0  | Disk         | 10 | 3        | 2            | 2                     | 4           | 2          | 0.000                          | 0.000                   | 1                    | 0                                     | 3                                         |
| G9    | 200, 0, 120  | Disk         | 11 | 3        | 3            | 0                     | 3           | 1          | 0.000                          | 0.000                   | 1                    | 0                                     | 3                                         |
| G10   | 77, 0, 96    | Disk         | 12 | 3        | 3            | 0                     | 3           | 3          | Applica                        | Applica                 | 3                    | 3                                     | 1                                         |
| G11   | 91, 0, 191   | Disk         | 13 | 2        | 2            | 0                     | 2           | 1          | 0.000                          | 0.000                   | 1                    | 0                                     | 2                                         |
| G12   | 0, 98, 130   | Disk         | 14 | 2        | 2            | 0                     | 2           | 1          | 0.000                          | 0.000                   | 1                    | 0                                     | 2                                         |
| G13   | 0, 12, 96    | Solid Square | 15 | 2        | 2            | 0                     | 2           | 1          | 0.000                          | 0.000                   | 1                    | 0                                     | 2                                         |
| G14   | 0, 136, 227  | Solid Square | 16 | 2        | 2            | 0                     | 2           | 1          | 0.000                          | 0.000                   | 1                    | 0                                     | 2                                         |
| G15   | 0, 100, 50   | Solid Square | 17 | 2        | 1            | 0                     | 1           | 0          | 0.000                          | 0.000                   | 1                    | 0                                     | 2                                         |



| S. No | Movie Name                    | Twitter handle        | Number of connected components |
|-------|-------------------------------|-----------------------|--------------------------------|
| 1     | Sweetie Weds NRI              | sweetiewedsnri        | 23                             |
| 2     | Mirror Game                   | mirrorgamefilm        | 23                             |
| 3     | Hanuman da Damdaaar           | hanumandadamdaar      | 104                            |
| 4     | Flat 211                      | flat211               | 14                             |
| 5     | Dobaara                       | dobaara               | 462                            |
| 6     | Dear Maya                     | dearmaya              | 198                            |
| 7     | Behen Hogi Teri               | behenhogiteri         | 252                            |
| 8     | Raabta                        | raabta                | 1250                           |
| 9     | Love You Family               | loveyoufamily         | 10                             |
| 10    | Bank Chor                     | bankchor              | 305                            |
| 11    | Tubelight                     | tubelight             | 924                            |
| 12    | Ek Haseena Thi Ek Deewana Tha | ehtedt                | 23                             |
| 13    | Mom                           | mommovie              | 30                             |
| 14    | Guest In London               | guestinlondon         | 97                             |
| 15    | Jagga Jasoos                  | jaggajasoos           | 672                            |
| 16    | Shab                          | shabthefilm           | 25                             |
| 17    | Lipstick Under My Burkha      | lipstickundermyburkha | 874                            |
| 18    | Munna Michael                 | munnamichael          | 427                            |
| 19    | Raag Desh                     | raagdes               | 216                            |
| 20    | Indu Sarkar                   | indusarkar            | 653                            |
| 21    | Mubarakan                     | mubarakan             | 780                            |
| 22    | Baat Company                  | baatcompany           | 15                             |

The number of communities in a graph also depends on the total number of vertices in that graph. In this case, each vertex is a Twitter ID whose associated user tweeted using the Twitter handle of the movie. Lesser number of vertices lead to lesser number of communities and hence a sparse network is generated. Movies which failed to capture the attention of the general public got a lesser number of tweets and the spread of information in the initial stage was slower and restricted by the number of users who initiated the opinion sharing process. The spiral graphs generated by NodeXL give a clear comparison of movie popularity by highlighting the number of edges and number of connected components.



**Fig. 2 Spiral Graphs of the entire Bollywood movie dataset**

Figure 2 shows the spiral graphs of all the summary files containing 7 day tweet datasets for the movies released during the test period. More the number of connected components, more is the number of communities in the network and hence a denser graph with vertices establishing complex relationships amongst themselves.

## V. CONCLUSION

As social networks are becoming more and more popular, there is a need to mine them and analyze the datasets from a new perspective to yield fruitful results. The Bollywood dataset analyzed in this research has presented methods of detection of communities in a newly released movie dataset which can help the personals related to this business to take help of social network analysis for their movie's promotion and understanding the spread of online word of mouth publicity through these communities.

## REFERENCES

- [1] J. Yang, J. Leskovec, Defining and evaluating network communities based on ground-truth, *Knowl. Inf. Syst.* 42 (2015) 181–213. doi:10.1007/s10115-013-0693-z.
- [2] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, *Proc. Natl. Acad. Sci.* 99 (2002) 7821–7826. doi:10.1073/pnas.122653799.
- [3] M. Ouyang, L. Zhao, L. Hong, Z. Pan, Comparisons of complex network based models and real train flow model to analyze Chinese railway vulnerability, *Reliab. Eng. Syst. Saf.* 123 (2014) 38–46. doi:10.1016/j.res.2013.10.003.
- [4] T. Tsekeris, Interregional trade network analysis for road freight transport in Greece, *Transp. Res. Part E Logist. Transp. Rev.* 85 (2016) 132–148. doi:10.1016/j.tre.2015.11.005.
- [5] R. Akerkar, P. Maret, L. Vercouter, Web intelligence and communities, *Proc. 4th Int. Work. Web Intell. Communities - WI&C '12.* (2012) 1. doi:10.1145/2189736.2189738.
- [6] P. Sah, L.O. Singh, A. Clauset, S. Bansal, Exploring community structure in biological networks with random graphs, *BMC Bioinformatics.* 15 (2014) 220. doi:10.1186/1471-2105-15-220.
- [7] M. Ley, *DBLP — Some Lessons Learned \**, (2009).
- [8] M. Weiss, H. Hultsch, I. Adam, C. Scharff, S. Kipper, The use of network analysis to study complex animal communication systems: a study on nightingale song, *Proc. R. Soc. B Biol. Sci.* 281 (2014) 20140460–20140460. doi:10.1098/rspb.2014.0460.
- [9] J. Leskovec, K.J. Lang, A. Dasgupta, M.W. Mahoney, Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters, *Internet Math.* 6 (2009) 29–123. doi:10.1080/15427951.2009.10129177.
- [10] S. Zannettou, T. Caulfield, E. De Cristofaro, N. Kourtellis, I. Leontiadis, M. Sirivianos, G. Stringhini, J. Blackburn, The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources, (2017). <http://arxiv.org/abs/1705.06947>.
- [11] J. Kleinfield, The small-world problem, *Soc. Sci. Public Policy.* 39 (2002) 61–66. doi:10.1007/BF02717530.
- [12] S. Fortunato, D. Hric, Community detection in networks: A user guide, (2016) 1–43. doi:10.1016/j.physrep.2016.09.002.
- [13] S. Fortunato, Community detection in graphs, *Phys. Rep.* 486 (2010) 75–174. doi:10.1016/j.physrep.2009.11.002.
- [14] D. Shizuka, D.R. Farine, Measuring the robustness of network community structure using assortativity, *Anim. Behav.* 112 (2016) 237–246. doi:10.1016/j.anbehav.2015.12.007.
- [15] J. Yang, J. McAuley, J. Leskovec, Community detection in networks with node attributes, *Proc. - IEEE Int. Conf. Data Mining, ICDM.* (2013) 1151–1156. doi:10.1109/ICDM.2013.167.
- [16] H. Situngkir, M P RA Spread of hoax in Social Media Spread of hoax in Social Media, (2011). <https://mpira.ub.uni-muenchen.de/30674/>.
- [17] D. Shah, T. Zaman, Rumors in a Network : Who ' s the Culprit ?, *Preprint.* 57 (2010) 1–43. doi:10.1109/TIT.2011.2158885.

- [18] S. Moon, P.K. Bergey, D. Iacobucci, Dynamic Effects Among Movie Ratings, Movie Revenues, and Viewer Satisfaction, *J. Mark.* 74 (2010) 108–121. doi:10.1509/jmkg.74.1.108.
- [19] S. Dooms, T. De Pessemier, L. Martens, MovieTweatings : a Movie Rating Dataset Collected From Twitter, *Work. Crowdsourcing Hum. Comput. Recomm. Syst. CrowdRec RecSys.* 2013 (2013) 43–44.
- [20] R. Sandeep, S. Sood, Exploring Twitter for Large Data Analysis, 6 (2016) 325–330.